



United States
Department
of Agriculture

Forest Service

**Rocky Mountain
Research Station**

General Technical Report
RMRS-GTR-216WWW

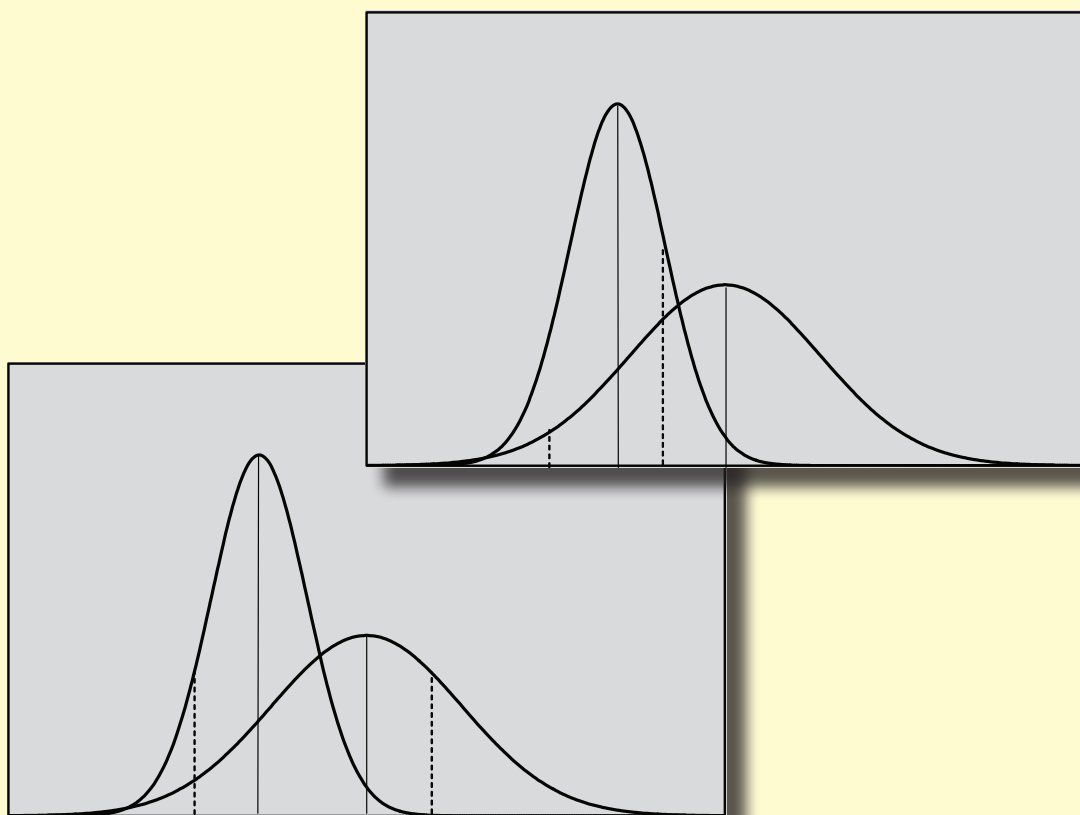
January 2009



An Enquiry Into the Method of Paired Comparison:

Reliability, Scaling, and Thurstone's Law of Comparative Judgment

Thomas C. Brown and George L. Peterson



Brown, Thomas C.; Peterson, George L. 2009. **An enquiry into the method of paired comparison: reliability, scaling, and Thurstone's Law of Comparative Judgment.** Gen Tech. Rep. RMRS-GTR-216WWW. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 98 p.

Abstract

The method of paired comparisons is used to measure individuals' preference orderings of items presented to them as discrete binary choices. This paper reviews the theory and application of the paired comparison method, describes a new computer program available for eliciting the choices, and presents an analysis of methods for scaling paired choice data to estimate an interval scale measure of preference. A new procedure for isolating an individual's inconsistent choices is described. Using data from five empirical studies, the reliability of respondents' paired choices is assessed using measures of internal reliability, choice consistency, and test-retest reliability.

Keywords: Public judgments, random utility, reliability, consistency, scaling, response time, preference learning

The Authors

Thomas Brown is an economist with the U.S. Forest Service, Rocky Mountain Research Station, Fort Collins, CO.

George Peterson retired in 2004 after 22 years with the U.S. Forest Service, Rocky Mountain Research Station, Fort Collins, CO. Previously, he was a professor at Northwestern University where he taught for 18 years.

Rocky Mountain Research Station Publishing Services

Telephone	(970) 498-1392
FAX	(970) 498-1122
E-mail	rschneider@fs.fed.us
Web site	http://www.fs.fed.us/rm/publications
Mailing Address	Publications Distribution Rocky Mountain Research Station 240 West Prospect Road Fort Collins, CO 80526

Acknowledgments

We are deeply indebted to Glen Brink, formerly with the Rocky Mountain Research Station, who did the programming on the two computer programs described herein. We also thank Randy Rosenberger of Oregon State University, Paul Bell of Colorado State University, and Andrea Clarke of the USDA's Natural Resources Conservation Service for helpful comments. And we thank Andrej Birjulin and Andrea Clarke for the use of data they collected while they were PhD students at Colorado State University.

Preface

A key feature of our research is a computer program we developed to administer a paired comparison survey to a respondent. This program includes a procedure to quickly select the subset of the respondent's choices that are inconsistent with the overall ordering implied by the full set of choices and then present the inconsistent choices a second time. Two versions of this program are mentioned herein. The initial version of the program was used in five paired comparison experiments that provide the essential data for this study. Based on what we learned from the implementation and testing of the initial version, we then developed a new version of the program. The improved program, but not the initial version, is available for others to use and is described in Chapter 2.

Contents

Chapter 1. Introduction	1
Paired Comparison Basics	2
An Early Random Utility Model: The Law of Comparative Judgment ..	5
Reliability	8
Consistency	8
Internal Reliability	8
Test-Retest Reliability	10
Scale Values	10
Aggregate Preference Scores	10
Psychometric Scaling	11
Summary	13
 Chapter 2. Methods	15
Paired Comparison (PC) Program	15
Input Files	15
Output Files	19
Paired Comparison Simulation (PCS) Program	21
PCS Program Modules	22
Empirical Studies	24
Statistical Procedures	29
Summary and Conclusions	29
 Chapter 3. Scaling	30
Scaling Procedures	30
IID Discriminal Dispersions	32
Evenly Spaced V s	32
Randomly Spaced V s	35
V s in Three Clusters	38
Heterogeneous Discriminal Dispersions	42
Evenly Spaced V s	42
Randomly Spaced V s	43
V s in Three Clusters	43
V s in Two Clusters	43
Gross Heterogeneity of V s and Discriminal Dispersions	44
Effect of Sample Size	45
Summary and Conclusions	46

Chapter 4. Effectiveness of the Double-Sort Procedure	48
Introduction	48
The DS Procedure	48
Simulation Results	50
Issues Regarding the Setup of the Paired Comparison	
Experiment	52
How Well Does the DS Procedure Work?	53
Summary	57
Evidence From Real-People Experiments	59
Conclusions	64
 Chapter 5. Consistency and Reliability	66
Introduction	66
Relationships Between Consistency and Internal Reliability	66
Individual Respondent Variation	66
Effect of Δ and Number of Items	67
Effect of Range of V	70
Effect of Clustered V s	71
Test-Retest Reliability	72
Evidence of Learning and/or Memory	74
Does Choice Difficulty Influence Consistency and Choice Switching?	78
Decision Time	78
Type of Item	79
Effect of Retest-Corrected Choices on Internal Reliability	80
Summary	81
 Chapter 6. Closing Thoughts	83
Literature Cited	86
Appendix: RUM and the Law of Comparative Judgment	88
The Psychologists' Random Utility Model	88
The RUM Model	89
A Reconciliation of the Two Random Utility Models	89
Frequently Used Abbreviations and Symbols	91
Endnotes	92

An Enquiry Into the Method of Paired Comparison:

Reliability, Scaling, and Thurstone's Law of Comparative Judgment

Thomas C. Brown and George L. Peterson

Chapter 1. Introduction

The method of paired comparisons for recording human judgments has a long history, dating back at least to the late 19th century.¹ An early English text on experimental psychology by Titchener (1901) covered paired comparisons, and Thurstone (1927a, 1927c) brought considerable attention to the method with his psychological scaling proposals in the late 1920's. Summaries of the method are found in psychometric textbooks (for example, Guilford 1954; Nunnally 1976; Torgerson 1958) and, from the statistical perspective, in the books by David (1988) and Bock and Jones (1968).

A paired comparison is simply a binary choice. With the method of paired comparisons, a set of stimuli, or items, is judged,² usually by presenting all possible pairs of the items to each respondent who chooses for each pair the item that better satisfies the specified choice criterion (for example, more preferred, more serious, more beautiful).³

Fechner (1860) used binary choices to study psychophysical relationships, that is, judgments of physical phenomena such as the weight of objects, brightness of lights, and loudness of sounds. For example, when presented with pairs of noises, respondents were asked in each case to say which was louder. Later, the method was extended to obtain value judgments, such as of the pleasantness of different colors (Titchener 1901), seriousness of crimes (Thurstone 1927b), scenic beauty of forest scenes (Buhyoff and Leuschner 1978), and seriousness of environmental losses (Brown and others 2002).

Advantages of paired comparisons as a method for eliciting human judgments include the method's simplicity and its use of comparative judgments. Regarding comparative responses, Titchener (1901, p. 92) notes: "We have no absolute measure of the amount of pleasantness or unpleasantness that corresponds to a given stimulus; but we are able to say, when two stimuli are presented, which of them is the more or which the less pleasant." Similarly, Nunnally (1976, p. 40) asserts that "People simply are not accustomed to making absolute judgments in daily

life, since most judgments are inherently comparative...people are notoriously inaccurate when judging the absolute magnitude of stimuli.” The method of paired comparisons thus uses our inherent familiarity with, and ability to, make comparisons. Among the methods available for eliciting comparative judgments, which in addition to binary choices include choices among more than two stimuli and ranking of several stimuli, binary choice is clearly the easiest for respondents because it compares only two stimuli at a time.

Our objective in this monograph is to provide a review and analysis of the use of the method of paired comparisons to assess public preferences. We will describe the method (Chapter 1), assess the utility of several ways of estimating scale values from respondents’ choices (Chapter 3), evaluate a new method for estimating the consistency of an individual choice with the respondent’s full set of paired choices (Chapter 4), and examine the reliability of a set of paired choices (Chapter 5). The results we present come from a simulation model we developed for assessing the ability of paired comparisons to accurately assess preferences and from several empirical studies using the method, all of which are described in Chapter 2.

Paired Comparison Basics

With paired comparisons, respondents choose the stimulus, or item, in each pair that has the greater magnitude on the choice dimension they were instructed to use. The simplest approach, which we will use throughout, is to present all possible pairs of the items to each respondent. With t items, there are $t(t - 1)/2$ pairs in total. Each pair results in a binary choice. The choices allow calculation of a set of *scale values* indicating the position of the items along the specified dimension. Scale values may be estimated for an individual respondent or for a group of respondents.

The choice dimension may be whatever fits the items being evaluated, including, as mentioned above, everything from physical dimensions to sociological judgments. For valuation applications, the items are either all gains (for example, commercial goods or environmental improvements) or all losses (for example, personal injuries or environmental damages).

When presented with a pair of items, respondents are not offered an indifference option. This practice is supported by the theory of stochastic preference, wherein the probability of true indifference at any one moment is assumed to be very small. The practice also has the practical benefit of maximizing the amount of information potentially learned about the respondent’s preferences. Although the lack of an indifference option may sometimes force respondents to make what seem like unwarranted distinctions, this is not worrisome, because across many comparisons, indifference between two items will be revealed in the data as an equality of scale values.⁴

For each respondent, the full set of choices yields a *preference score* for each item, which is the number of times the respondent preferred the item to other items in the set. Preference scores are the simplest form of scale values for the items. These scores are easily calculated by creating a t by t matrix and entering a 1 in each cell where the column item was preferred to the row item, or a 0 otherwise.

Column sums give the preference scores. For example, figure 1-1 contains a hypothetical matrix for a 10-item choice set. Preference scores, at the bottom of the matrix, indicate that, for example, item 5, with a preference score of 7, was chosen seven of the nine times it appeared among the choices.

Mathematically stated, the preference score (in other words, the column sum) of an item is:

$$a_c = \sum_{r=1}^t a_{rc} \quad (1)$$

where a_{rc} is the cell score (either 0 or 1) for a given row (r) and column (c). The row sum is:

$$a_r = \sum_{c=1}^t a_{rc} = (t-1) - a_c \quad (2)$$

As indicated in equation 2, row sums are a mirror image of column sums.

A respondent's vector of preference scores describes the individual's preference order among the items in the choice set, with larger integers indicating more preferred items. The number of items in the set determines the preference score range, which is always from 0 to $t-1$. So, for 10 items, the range is from 0 to 9.

a.) Raw choice matrix

item	1	2	3	4	5	6	7	8	9	10	sum
1	--	0	0	1	1	1	1	1	1	1	7
2	1	--	1	1	1	1	1	1	1	1	9
3	1	0	--	1	1	1	1	1	1	1	8
4	0	0	0	--	0	0	0	0	1	0	1
5	0	0	0	1	--	0	0	0	1	0	2
6	0	0	0	1	1	--	1	0	1	1	5
7	0	0	0	1	1	0	--	0	1	1	4
8	0	0	0	1	1	1	1	--	1	1	6
9	0	0	0	0	0	0	0	0	--	0	0
10	0	0	0	1	1	0	0	0	1	--	3
sum:	2	0	1	8	7	4	5	3	9	6	

b.) Double-sorted choice matrix

item	2	3	1	8	6	7	10	5	4	9	sum
2	--	1	1	1	1	1	1	1	1	1	9
3	0	--	1	1	1	1	1	1	1	1	8
1	0	0	1	1	1	1	1	1	1	1	7
8	0	0	0	--	1	1	1	1	1	1	6
6	0	0	0	0	--	1	1	1	1	1	5
7	0	0	0	0	0	--	1	1	1	1	4
10	0	0	0	0	0	0	--	1	1	1	3
5	0	0	0	0	0	0	0	--	1	1	2
4	0	0	0	0	0	0	0	0	--	1	1
9	0	0	0	0	0	0	0	0	0	--	0
sum:	0	1	2	3	4	5	6	7	8	9	

Figure 1-1. A perfectly consistent individual choice matrix for 10 items ($C = 0$, $\zeta = 1.0$).

For a given respondent, the difference between the preference scores of items in a pair is the *preference score difference* (PSD). This integer can range from 0 to $t-1$.

A *circular triad* is an intransitive preference order among three items. For example, three binary choices among items i, j , and k may produce the following order: $k > j > i > k$. It is not, of course, clear which of the three choices caused the circular triad. Any one of the three could be the culprit. If an individual respondent is sampled only once, a lack of internal reliability is detectable only if it causes one or more circular triads.

In the case of a 10-item choice set, an individual's preference score vector with no circular triads contains all 10 integers from 0 through 9 (fig. 1-1a). Sorting the matrix by column sum, increasing from left to right, arranges the items in order of increasing preference score. Also sorting the matrix by row sum, increasing from bottom to top, produces a *double-sorted* matrix (fig. 1-1b). If the double-sort matrix contains no circular triads it will have a "1" in all the cells above (and thus to the right of) the principal diagonal and a "0" in all cells below the principal diagonal.

Figure 1-2 shows a 10-item data set identical to that of figure 1-1 except that the choice between items 3 and 7 has been reversed (item 3 was chosen instead of item 7, whereas in fig. 1-1, item 7 was chosen). This reversal causes three circular triads.

a.) Raw choice matrix.

item	1	2	3	4	5	6	7	8	9	10	sum
1	--	0	0	1	1	1	1	1	1	1	7
2	1	--	1	1	1	1	1	1	1	1	9
3	1	0	--	1	1	1	0	1	1	1	7
4	0	0	0	--	0	0	0	0	1	0	1
5	0	0	0	1	--	0	0	0	1	0	2
6	0	0	0	1	1	--	1	0	1	1	5
7	0	0	1	1	1	0	--	0	1	1	5
8	0	0	0	1	1	1	1	--	1	1	6
9	0	0	0	0	0	0	0	0	--	0	0
10	0	0	0	1	1	0	0	0	1	--	3
sum:	2	0	2	8	7	4	4	3	9	6	

b.) Double sorted matrix.

item	2	1	3	8	6	7	10	5	4	9	sum
2	--	1	1	1	1	1	1	1	1	1	9
1	0	--	0	1	1	1	1	1	1	1	7
3	0	1	--	1	1	0	1	1	1	1	7
8	0	0	0	--	1	1	1	1	1	1	6
6	0	0	0	0	--	1	1	1	1	1	5
7	0	0	1	0	0	--	1	1	1	1	5
10	0	0	0	0	0	0	--	1	1	1	3
5	0	0	0	0	0	0	0	--	1	1	2
4	0	0	0	0	0	0	0	0	--	1	1
9	0	0	0	0	0	0	0	0	0	--	0
sum:	0	2	2	3	4	4	6	7	8	9	

Figure 1-2. An individual choice matrix with three circular triads ($C = 3$, $\zeta = 0.93$).

Circular triads cause some integers to appear more than once in the vector of preference scores, as do preference scores of 2 and 4 in figure 1-2, while others disappear. In general, the greater the PSD of the items involved in the inconsistent response, the greater the number of circular triads that result. Circular triads cause “1”s to appear below the principal diagonal, and thus “0”s above the principal diagonal, of the double-sorted matrix (fig. 1-2b). The double-sorted matrix and its relation to circular triads are described in more detail in Chapter 4.

An Early Random Utility Model: The Law of Comparative Judgment

Inconsistency among binary choices was studied by psychologists in the 19th century. They found that people’s comparative judgments of physical stimuli (such as weight of objects) became less accurate the more alike were the stimuli. For example, the closer the weights of paired objects, the greater the proportion of subjects who misjudged which object of a pair was heaviest, and the greater the likelihood that a given subject would misjudge the relative weights some of the time in repeated trials (Guilford 1954). Also, differences in repeatability of binary judgment were later observed with qualitative judgments, such as the excellence of hand writing samples or the seriousness of offenses (Thurstone 1927b). Subjects had more difficulty consistently judging some pairs than others. It was assumed that, as with physical stimuli, increasing inconsistency was associated with increasing similarity of the items on the dimension of interest.

In an effort to explain inconsistent paired comparison responses, Thurstone (1927a) proposed a model characterizing judgment as a stochastic or random process, wherein an item falls along a *discriminal dispersion* around the modal value for the item on the *psychological continuum*. The dispersion was attributed to random errors in judgment. This model was soon applied to preferences, with the dispersion attributed to random fluctuations in preference. The model relies most importantly on the simple observation that the error in comparative judgments, as observed in the psychophysical studies mentioned above, increased as the two items being compared become closer along the psychological continuum (which we will call the *preference continuum*).

Thurstone (1927a) called his theory of binary choice the *law of comparative judgment*. According to that theory, an item has, for any given respondent, a fixed expected value on a continuum defined by the choice criterion. The respondent acts as if randomly sampling the value of the stimulus from a discriminial dispersion about that expected value. A binary choice between two items samples the instantaneously perceived difference in the values of the two items. The item chosen is the one with the largest perceived value. This psychometric theory is similar to the “constant utility” model described by Ben-Akiva and Lerman (1985) and attributed to Luce and Suppes (1965) in which the utilities of the items are constant.⁵ In Thurstone’s model it is assumed that those constant utilities are not single quantities (or intrinsic properties of the items), but rather are the expected values of the decision maker’s discriminial dispersions.

The characterization of preference as a stochastic process is formalized as a value function (U) consisting of systematic (deterministic) and random (error) components. For example, the value of item i to person p can be represented by the following relation:

$$U_{ip} = V_{ip} + \varepsilon_{ip} \quad (3)$$

where U is a momentary relative magnitude internal to the person. The systematic component, V , represented, for Thurstone, the expected value of U and the error component, ε , represented momentary variability about the expected value due to unobservable influences within the respondent. In the absence of strategic behavior, items of higher V will tend to be chosen above items of lower V in a paired comparison exercise. However, because of the randomness inherent in perception or judgment, responses may not always match the order of the expected values of items. Consider the value of two items i and j :

$$\begin{aligned} U_{ip} &= V_{ip} + \varepsilon_{ip} \\ U_{jp} &= V_{jp} + \varepsilon_{jp} \end{aligned} \quad (4)$$

If the error distributions of the items (ε_{ip} and ε_{jp}) overlap, the order of the values of the items at a given moment (U_{ip} and U_{jp}), and therefore the response, may be inconsistent with their respective expected values (V_{ip} and V_{jp}). The probability (π) that item j will be considered to be of greater value than item i is:

$$\pi(U_{jp} > U_{ip}) = \pi(V_{jp} + \varepsilon_{jp} > V_{ip} + \varepsilon_{ip}) \quad (5)$$

This probability increases as the difference ($V_{jp} - V_{ip}$) becomes larger or the error distributions (ε_{ip} and ε_{jp}) become narrower.

Thurstone proposed letting U_i and U_j be normally distributed. With independently and normally distributed errors, given expected values V_i and V_j and variances Δ_i^2 and Δ_j^2 , ($U_j - U_i$) is also normally distributed with:⁶

$$\begin{aligned} E(U_j - U_i) &= V_j - V_i \\ \Delta_{(U_i + U_j)}^2 &= \Delta_i^2 + \Delta_j^2 \end{aligned} \quad (6)$$

Figure 1-3 depicts the value of two items along the preference continuum, assuming the errors are normally distributed. Given the preferences of figure 1-3, responses will usually indicate that item j is preferred to item i . However, as indicated by the overlapping error distributions of items i and j , item i will sometimes be preferred to item j . A preference for item i over item j , although perfectly reasonable given the value function of figure 1-3, would be inconsistent with the expected values of the preference distributions. A successful method of eliciting human value judgments will estimate the expected values despite the variability of the response process.

As noted by Dillon and others (1993), McFadden (2001), and others, Thurstone's law of comparative judgment laid the foundation for the random utility maximization (RUM) model. It can be shown that the RUM model formalized by Manski (1977) is consistent with the psychometric model if we assume that the RUM is maximizing expected utility, not instantaneous utility.

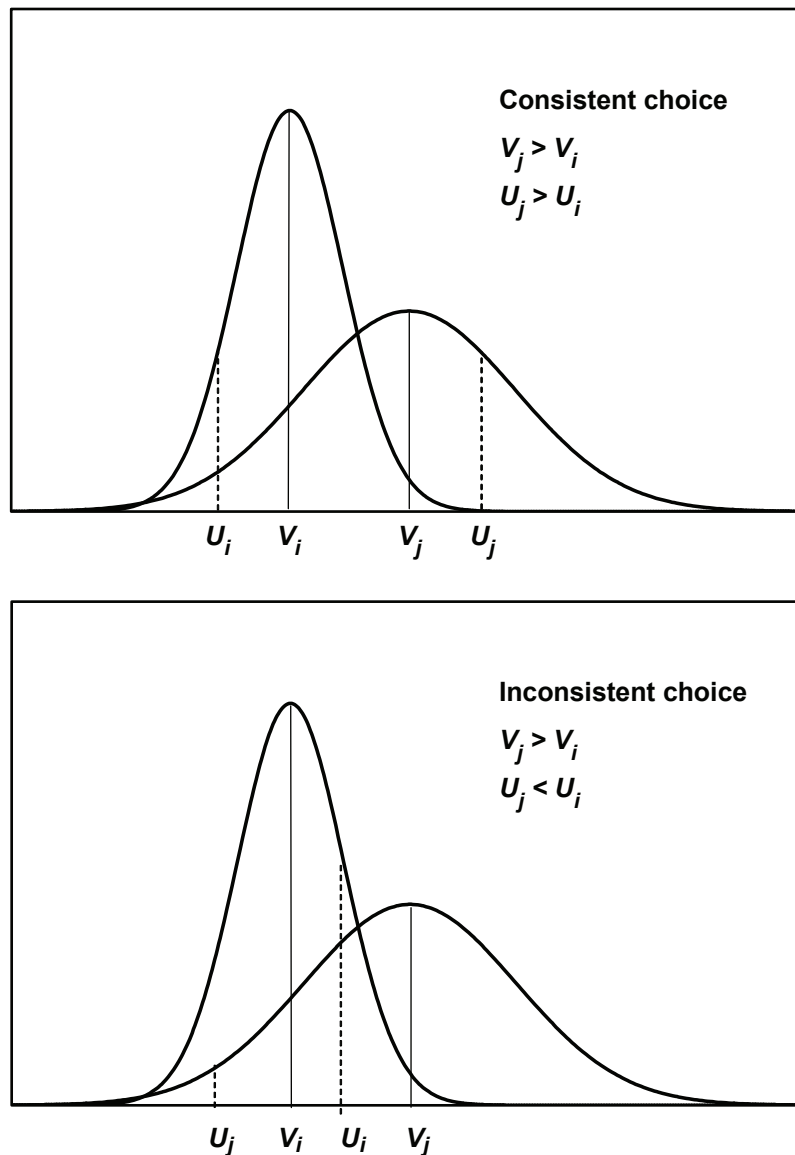


Figure 1-3. The law of comparative judgment.

The RUM model is notationally identical to Thurstone's model, but as commonly employed, the sources of error in the two models are quite different. Although Thurstone acknowledged the possibility of error both within the individual respondent and on the part of the researcher, in common application in psychological studies the error is within the individual respondent. However, with the modern RUM model, the error is in the researcher's ability to measure and model the choice process. The difference reflects two things: the data and the underlying models of human choice behavior. The law of comparative judgment was developed within psychology, where detailed studies of individual people had a long tradition and where error in individual judgment had been observed and studied for many decades. The RUM model developed in economics and related fields where the prevailing model of human choice behavior posited that individuals choose so as to maximize their utility and the data of interest tended to include only one choice per individual.

The law of comparative judgment was developed in an attempt to explain and model an individual person's multiple choices in carefully controlled experiments where uncontrolled (unmeasured) explanatory variables were unlikely. The RUM model was developed to explain and model multiple persons' single choices, where those choices were typically made in real life situations (for example, choice of a transportation mode) where uncontrolled explanatory variables were likely. See the Appendix for more on how these models compare.

Reliability

We define reliability in terms of three concepts. *Consistency* measures the respondent's ability to make binary choices that are consistent with the expected values of the items in each pair. *Internal reliability* questions whether a respondent's binary choices between items taken in pairs from a set of items are in agreement with each other, that is, whether they produce intransitivity as manifest by circular triads. *Test-retest reliability*, the usual meaning of reliability, asks about the degree of agreement between choices for identical pairs presented at different times. Consistency is a theoretical entity in that the expected values upon which this measure is based are non-observable. Internal reliability and test-retest reliability are empirical measures.

Consistency

Building on the prior discussion of the law of comparative judgment, we define consistency, and thus inconsistency, for an individual respondent as follows:

$$\text{Consistency} \quad U_j > U_i \text{ when } V_j > V_i \quad (7)$$

$$\begin{aligned} &U_j < U_i \text{ when } V_j < V_i \\ \text{Inconsistency} \quad &U_j < U_i \text{ when } V_j > V_i \\ &U_j > U_i \text{ when } V_j < V_i \end{aligned} \quad (8)$$

The probability that a paired comparison choice will be consistent, assuming $V_j > V_i$, is

$$\pi(\text{Consistency}) = \int_0^{\infty} f(U_j - U_i) \quad (9)$$

The probability that a choice will be inconsistent, again assuming $V_j > V_i$, is

$$\pi(\text{Inconsistency}) = \int_{-\infty}^0 f(U_j - U_i) = 1 - \pi(\text{Consistency}) \quad (10)$$

Although consistency cannot be measured directly, it can be modeled, and it can be observed indirectly, as described in Chapter 4.

Internal Reliability

Internal reliability refers to the level of agreement among a given respondent's choices and is thus a between-choice measure. A common measure of internal reliability is based on the number of circular triads among a respondent's choices.

The causes of circular triads in a paired comparison experiment include (1) respondent inability to consistently discriminate between similar items, (2) dominance of different attributes in different pairs, (3) order effects, and (4) respondent carelessness (a mistake in indicating a choice), incompetence (lack of understanding of the task), or intentional misrepresentation of preferences. The first cause results from the existence of positive error terms (ϵ), whereas the second and third causes involve changes in the V s over the course of the paired comparison survey.⁷ The second cause of circular triads can only occur with multi-attribute items, and then only if respondents focus on different attributes in different comparisons.⁸ The third cause of circular triads occurs when the V s depend at least partly on which pairs of items appeared earlier.

In the empirical and simulated examinations of reliability to follow, we assume that the V s remain constant for an individual respondent. Thus, for the purpose of this discussion we assume that all observed circular triads are randomly generated (in other words, that they result from the instantaneous random draws from the preference distributions about the V s), that the expected outcome complies with the transitivity axiom of rational utility theory, and, therefore, that the expected values satisfy the IIA (independence from irrelevant alternatives) property, as argued by Luce (1959).⁹

We must note, however, that homogeneity among real respondents in an empirical paired comparison experiment is unlikely. That is, there is no reason to expect that compliance with the IIA condition at the individual level will ensure compliance across an aggregation of multiple respondents (Ben-Akiva and Lerman 1985, p. 109-111). We must therefore assume—as in empirical applications of the random utility model¹⁰—that the combined effect of the dispersion of preferences among respondents and the discriminial dispersions within respondents acts, in effect, as if the sample were homogeneous.

The *coefficient of internal reliability*, ζ , was defined by Kendall and Smith (1940) as:

$$\zeta = 1 - \frac{C}{C_{\max}} \quad (11)$$

where C is the number of circular triads in the matrix of binary choices and $C_{\max}$ is the maximum possible number of circular triads.¹¹ This measure requires that all possible pairs of items in the set have been judged and varies between 1 for no circular triads and 0 for the maximum possible number. The number of circular triads in an individual's responses can be calculated directly from the preference scores. Letting t equal the number of items in the set, a_i equal the preference score of item i , and b equal the average preference score (in other words, $(t-1)/2$), the number of circular triads for an individual respondent, C , is:

$$C = \frac{t}{24}(t^2 - 1) - \frac{1}{2} \sum (a_i - b)^2. \quad (12)$$

Kendall and Smith derive the coefficient of internal reliability by proving that the maximum possible number of circular triads is:¹²

$$\begin{aligned} C_{\max} &= \frac{t(t^2 - 1)}{24}, & t \text{ odd}, \\ C_{\max} &= \frac{t(t^2 - 4)}{24}, & t \text{ even}. \end{aligned} \tag{13}$$

Perfect consistency produces perfect internal reliability. As will be seen in later chapters, inconsistency usually, but not always, produces lack of internal reliability.

Test-Retest Reliability

A short-term test-retest measure of reliability is made possible by repeating a random selection of pairs at the end of a paired comparison session. Some switching of choice over multiple presentations of the same pair is expected because of the stochastic nature of preference, whereby the values (U) of items change slightly from one moment to the next, as in equation 3. Switching is more likely when the V s of the items of the pair are close to each other and the $\mathcal{E}$ s are large. The closer the V s of the items, the smaller the PSD, all else equal.

Test-retest reliability is related to consistency. Assuming the choices are independent, the probability that a respondent will reverse an inconsistent choice in a second trial is

$$\pi(\text{Reverse}|\text{Inconsistency}) = \pi(\text{Consistency}) \tag{14}$$

and the probability that a consistent choice will be reversed in a second trial is

$$\pi(\text{Reverse}|\text{Consistency}) = 1 - \pi(\text{Consistency}) \tag{15}$$

When the choice set is large, we suggest that the assumption that choices are independent is reasonable for most respondents in the sense that respondents generally do not remember their prior choices.¹³

Scale Values

Aggregate Preference Scores

The response matrices of all respondents in the sample can be summed to provide an aggregate choice matrix, also called a frequency matrix, for the sample. For example, figure 1-4 shows a hypothetical frequency matrix for a sample of 10 respondents who each judged all possible pairs of 10 items. Column sums give the aggregate preference scores for the sample, specifying the number of times each item was chosen across all paired comparisons made by the respondents. Aggregate preference scores are converted to mean preference scores (MPSs) by dividing by the number of respondents.

item	1	2	3	4	5	6	7	8	9	10
1	--	0	0	9	10	10	10	10	10	10
2	10	--	10	10	10	10	10	10	10	10
3	10	0	--	9	7	9	1	10	10	10
4	1	0	1	--	1	2	1	1	10	0
5	0	0	3	9	--	1	1	0	10	0
6	0	0	1	8	9	--	10	0	8	10
7	0	0	9	9	9	0	--	0	10	10
8	0	0	0	9	10	10	10	--	10	10
9	0	0	0	0	0	2	0	0	--	0
10	0	0	0	10	10	0	0	0	10	--
sum:	21	0	24	73	66	44	43	31	88	60
MPS:	2.1	0.0	2.4	7.3	6.6	4.4	4.3	3.1	8.8	6.0
proportion:	0.23	0.00	0.27	0.81	0.73	0.49	0.48	0.34	0.98	0.67

Figure 1-4. An aggregate raw choice matrix for 10 respondents.

The bottom line in figure 1-4 illustrates another way to express the preference scores—as the proportion of times the item was chosen relative to the maximum possible number of times, which in this case is $10 \times 9 = 90$ times.

The aggregate preference scores, and thus the scale values, indicate the ordinal position of the items and approximate an interval scale measure of preference, revealing the sizes of the intervals between items. As will be seen in Chapter 3, the approximation, given a sufficient sample size, tends to be quite accurate except perhaps near the ends of the scale (in other words, except close to the most and least preferred items). The accuracy of the interval information is less near the ends of the scale because the paired comparison choice data are less rich there. That is, the lack of items beyond those of minimum and maximum V limits the number of choices that could help map out interval sizes in the regions of those items along the preference continuum.

Psychometric Scaling

Thurstone (1927a) (see also Bock and Jones 1968; Guilford 1954; Torgerson 1958) proposed a more involved scaling procedure in his “law of comparative judgment” that in theory corrects for the problem near the ends of the scale. His approach models $(V_{ip} - V_{jp})$ as a function of the probability of $(U_{ip} > U_{jp})$. The approach relies on the assumption that U is normally distributed about V , along with other assumptions regarding the variances of and correlations between the disturbance terms. Thurstone’s full model of comparative judgment allows for the variances of the error terms to vary from item to item and for the variances of the different items to be correlated. Having proposed the full model of comparative choice, Thurstone then offered a series of simplifying assumptions, the acceptance of which made computation of scale values from paired comparison data first of all feasible, and then, with additional assumptions, easier.¹⁴ Accepting the full set of assumptions, Thurstone’s Case V amounts to accepting what is now known as the IID condition (independently and identically distributed errors), wherein the variances of the error terms are uncorrelated (in other words, independent) and identical across all the items.

Independence among choices precludes order effects. Having a constant error variance, along with independence, greatly simplifies the derivation of scale values.¹⁵

Thurstone also suggested the possibility that the assumptions would apply not only across multiple choices of a single respondent but also across a whole set of respondents, wherein the variances are assumed to be identical not only across items but across individuals, producing one constant variance (Δ^2).¹⁶ If the V s of the various respondents are also assumed to be identical for a given item, providing a homogeneous sample of individuals, the paired comparison responses of a group of respondents can be combined to provide a set of scale values accurately representing the group.¹⁷

Assuming identical variances across items and respondents, Thurstone's Case V proposes that the scale value of an item equals the mean of the unit normal deviates corresponding to the probabilities that the item is chosen above each of the other items in the choice set. The probabilities are estimated by the proportions of times an item was chosen above other items in the paired comparison survey, which are easily computed from the frequency matrix.

We will not describe Thurstone's model in detail, as his approach is nicely described by others (for example, Bock and Jones 1968, Torgerson 1958). The central idea, however, can easily be summarized here. Consider the idealized graphs of figure 1-5, which depict identically distributed normal error distributions for three items assuming the V s are evenly spaced. The top graph shows two items that differ in V by one-half of Δ , with a third item several intervals to the right.

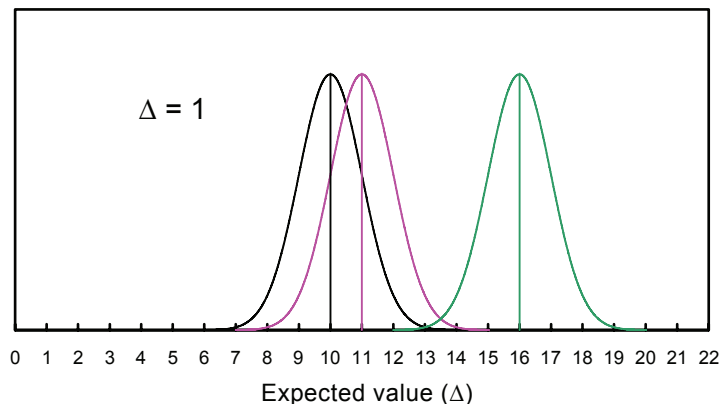
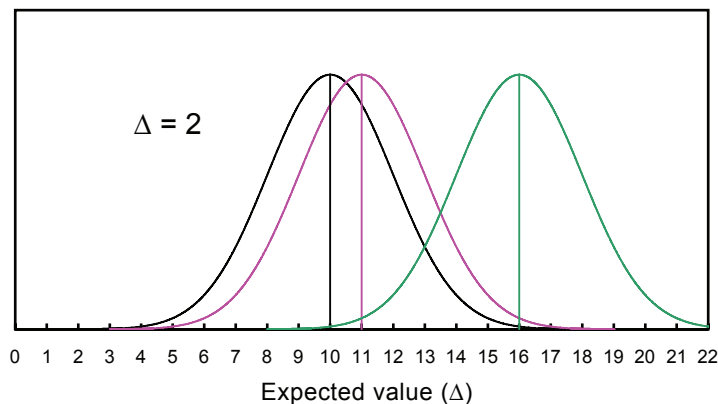


Figure 1-5. Locations of the error distributions of three items with V s of 10, 11, and 16, assuming identically distributed normal errors.

The ample overlap of the two distributions on the left suggests that inconsistent choices involving these items would not be uncommon. The bottom graph shows less overlap for the proximate distributions; thus, we would expect fewer inconsistent choices than shown in the top graph. Assuming that the dispersions are also independently distributed, the probability of an inconsistent choice can easily be computed using a table of areas below the normal curve of the difference distribution.¹⁸ Thurstone's approach takes advantage of this simple process, but reverses it. Knowing the probability that one item will be chosen over another (from a cell of the frequency matrix), and assuming a value for Δ , the difference in scale values for the pair of items can be estimated.

The independence assumption is anything but trivial. If this assumption is not met, the resulting preference scores or other scale values are dependent on the particular order in which the choices were presented. Multi-response valuation methods are potentially susceptible to two types of systematic changes that may occur over the course of a series of responses and that would violate the independence assumption: the expected values of the items may change and the error distributions may change. An effect of order on the expected values, to the extent it occurs, cannot be avoided at the individual respondent level, but it can be neutralized across the sample by randomizing for each respondent the order in which the pairs appear.

Regardless of the order in which pairs of items are presented, systematic changes in the error distributions may occur over the course of a series of responses. Two such changes seem plausible. First, fatigue may cause responses to become erratic, leading to enlarging error distributions (ϵ) and therefore increasing inconsistency over the course of the experiment. Second, the processing of multiple valuations may lead respondents to become more certain about their preferences, leading to decreasing ϵ and thus increasing consistency with sequence.¹⁹

Luce (1959), McFadden (1974), and others proposed another approach to analyzing binary response data. If Thurstone's assumption about the error term is replaced by one of independent double-exponential random disturbances, we have the basis of the logit model, as the difference distribution of two independent double-exponential random variables is the logistic distribution.²⁰ It is important to note, however, that Thurstone's approach and the logit approach each require data sets that comply with some rather demanding assumptions and contain enough cases to allow capture of the full extent of the variability of preference within the population. Without a rich data set, the improvements over aggregate preference scores promised by their methods may not be achieved.²¹

Summary

Paired comparisons offer a wealth of information on respondents' relative preferences for, or judgments about, a set of items. This information includes not only indications of where the items fall along the dimension of interest, but also information about the reliability of a respondent's choices. Because each respondent provides numerous choices involving each item, the internal reliability of each respondent's set of choices can be measured. In addition, if a measure

of the consistency of an individual choice (for example, consistency with the scale values established based on the full set of choices from the respondent) is available, the change in consistency over the sequence of choices or across different types of pairs of items is possible for each respondent.

Regarding measures of preference or judgment, simple arithmetic mean preference scores can be computed that indicate the relative placement of the items along the dimension of interest. At a minimum, these scores offer an ordinal relation among the items. A topic of primary interest is the degree to which the data also indicate the relative sizes of the intervals between the items. Because interval-level data support parametric statistics, interval-level measures are much preferred to simple ordinal relationships.

Mean preference scores are often accepted as having interval properties, but there is little theoretical support for that acceptance. In a search for more robust measures, researchers have proposed a number of approaches. In the psychometric literature, research over many years by many different scientists has demonstrated that people's ability to consistently judge pairs of stimuli is related to how close the stimuli are on the dimension of interest—the closer the items, the lower the ability of people to consistently compare them. This inconsistency turns out to be very useful, as it provides information about how close the stimuli are on the value continuum corresponding to the dimension of interest. Psychometric scaling methods, designed to make use of the inconsistency, theoretically offer improvements over simple mean preference scores.

In the following chapters we examine the reliability of paired comparison choices, explore a new method for isolating individual inconsistent choices from among a respondent's set of paired comparisons, and evaluate the ability of scaling methods—both mean preference scores and psychometric procedures—to represent the true values of items being judged.

Chapter 2. Methods

We developed two computer programs to facilitate the study of paired comparisons – a program that administers a set of paired comparisons to actual respondents (called the PC program) and a program to simulate the responses of virtual respondents. We and others used the PC program in five experiments summarized herein.²² Here we describe the two computer programs and the five paired comparison experiments.

Paired Comparison (PC) Program

The PC program administers the paired comparison experiment on a personal computer that presents the pairs of items on the monitor in a unique random order to control for order effects. The items appear side-by-side on the monitor, with their position (right versus left) also randomized. The respondent enters a choice by pressing the right or left arrow key and can correct a mistake by pressing the backspace key and then making a new selection. At the end of the initial paired comparisons, the program runs the double-sort procedure (DS) and then can repeat in random order those pairs for which the individual's choice was determined by the DS procedure to be inconsistent. The program can also randomly select 10 consistent pairs for retesting. The individual pairs in these two sets of repeated choices are randomly intermixed. There is no pause or other indication to the respondent that the original pairs have ended and the repeats have begun. The computer also records the time taken to enter each original choice.

The original PC program (PC.EXE) was written in FORTRAN. That version was used for the five experimental studies discussed herein. Since then, the PC program has been updated. Portions of the code have been converted to Visual Basic to facilitate use with the Windows operating system, and capabilities have been added, such as the option to include photographs depicting the items, the option to obtain rating judgments of the items, and the option to improve the performance of the DS procedure used to select inconsistent choices.

The current version of the PC program (PAIRCOMP6.EXE) allows for presentation of four sets of questions in the following order: (1) general rating scale questions, (2) paired comparisons, with or without retesting, (3) ratings of the items presented in the paired comparisons, and (4) demographic questions. The paired comparisons are required and the remaining three sets of questions are optional.

Input files

Use of the PC program requires only one user-generated file, PAIRCOMP.TXT, which provides instructions to PAIRCOMP6.EXE. Figure 2-1 contains an example PAIRCOMP.TXT file. As seen in the figure, PAIRCOMP.TXT contains instructions for its preparation in comment statements (those beginning with ###). The initial portions of the comment statements—the ### and the capitalized title—must be included in PAIRCOMP.TXT exactly as shown in figure 2-1. For operation, PAIRCOMP.TXT and PAIRCOMP6.EXE must be located in the same folder.

WELCOME SCREEN. 3 lines required, followed by a blank line. 1st line, maximum of approximately 32 characters; 2nd line, maximum of approximately 420 characters; 3rd line, approximately 64 characters. This paircomp.txt file is for use with PairComp6.exe, which allows for presentation of 4 sets of questions in the following order: (1) rating scale questions (optional); (2) paired comparisons, with or without retrials (required); (3) ratings of the items presented in the paired comparisons (optional); and (4) demographic questions (optional). Note that this order is not the same as the order in which these sets of questions are encountered below.

Welcome to PAIRCOMP

This is an interactive computer program that has been designed to make your participation easier, faster and fun. Please read all of the instructions and follow them carefully when answering questions.

Please wait for further instruction on how to begin.

INSTRUCTION SCREEN, PAIRS PRESENTATION. 3 lines required, followed by a blank line. 1st line goes into the 1st label caption; 2nd line goes into the 2nd label caption; 3rd line goes into the 3rd label caption. The 3rd line may be blank (note that if the 3rd line is blank, the final blank line is still required, so that there will be 2 blank lines).

You will be asked to choose an item from each of several pairs of items.

For each pair, please choose the item that you would prefer to have, if they were each available for free.

You can backup one screen to correct a mistake or change a selection by clicking "Previous Screen".

PAIR PRESENTATION SCREEN, INSTRUCTIONS. 2 lines required, followed by a blank line. 1st line gives abbreviated instructions, maximum of approximately 400 characters. 2nd line gives a button label, maximum of approximately 15 characters.

Choose the item you would prefer to have.

I choose this one

PAIR PRESENTATION SCREEN, ITEM DEFINITIONS. One line per item, each with 4 fields, is required, followed at the end by a blank line. Fields are separated by commas (no imbedded commas are allowed in fields, and all 3 separating commas must appear).

1st field, item code; 2nd field, either blank, or "X" to exclude pairs with other items marked "X"; 3rd field, item description, maximum of approximately 200 characters; 4th field, item image file name (may be blank).

C1,X,\$1,\$1.jpg

C2,,Big Truck,BigTruck.jpg

C3,,Small Boat,SmallBoat.jpg

C4,X,\$10,\$10.jpg

C5,,Flowers,flowers.jpg

DOUBLE SORT TIES BY. Next line must be either Maintaining input order OR Weighting choices

Maintaining input order

PRESENT RETRIALS? 1 line with 2 fields, followed by a blank line. 1st field is the threshold coefficient of internal reliability (aka coefficient of consistency), a value between (but not including) 0.0 and 1.0. 2nd field is the threshold preference score difference (PSD), an integer that may be 0. A comma must separate the 2 fields. If no retrials are to be presented, the line contains only the comma in column 1.
0.3,0

RANDOMIZED ORDER RATING SCALE SCREEN INSTRUCTIONS. Optional, but 2 lines are required anyway, followed by a blank line.

1st line, instructions, maximum of approximately 420 characters (leave this line blank if item rating questions are NOT to be presented);

2nd line, 2 fields, one for left end of scale, one for right, approximately 30 characters each (if rating questions are not to be presented, this line must still have at least a comma).

On the rating scale below, please indicate how desirable you find the item.

Not at all desirable,Extremely desirable

STATEMENT RATING SCALE SCREEN INSTRUCTIONS. Optional, but 2 lines are required anyway (if no statements will be presented, these 2 lines may be blank except for the comma in the second line, which must be present), followed by a blank line. 1st line, instructions, maximum of approximately 420 characters; 2nd line, rating scale endpoints.

Please answer the question using the rating scale below.

Not at all,Extremely

STATEMENTS. One line per statement, all followed by a blank line. If there are no statements to be presented, leave only the 1 blank line.

Are you interested in big trucks?

Are you interested in boats?

Do flowers make you happy?

DEMOGRAPHICS SCREEN, QUESTION INSTRUCTIONS. Optional, but 2 lines are required (although they may be blank if no questions will be presented), followed by a blank line. 1st line, instructions, maximum of approximately 420 characters. 2nd line, Title for Next button

Please tell us a little about yourself by typing your answer in the box below each of the following questions. When the options are numbered, please type in the appropriate number. Click "Next" to move on.

Next

QUESTIONS. One line per question, all followed by a blank line; if there are no questions to be presented, leave only the 1 blank line.

What is your gender? (1-Male or 2-Female)

What is your age?

GOODBYE SCREEN. 1 line required, maximum of approximately 420 characters.

This is the end of PAIRCOMP. Thank you very much for your participation! We appreciate your time and effort.

Figure 2-1. PC program input (PAIRCOMP.TXT), with example statements.

The items presented for paired comparisons may be anything that can be described, including goods, services, environmental conditions, personal losses, environmental losses, and monetary amounts. The number of items may be as low as two, with no realistic upper limit. All items are compared with all other items unless the user indicates otherwise.

The PC program allows for photographs to be presented above their verbal descriptions when the items are presented for paired comparisons and when they are individually presented for ratings. The photos are provided in individual files, such as .jpg or .bmp files.²³ If used, these files must be located in the same folder as PAIRCOMP.TXT and PAIRCOMP6.EXE.

PAIRCOMP.TXT contains 12 sections, each beginning with a comment (###) line that describes the information needed and how it must be formatted (fig. 2-1). The first three sections contain instructions to the respondent. The fourth section, ###, PAIR PRESENTATION SCREEN, ITEM DEFINITIONS, lists the items to be compared, one line per item. Each line begins with an item code. The first character of the item code specifies item type. Subsequent characters of the item code allow the user to uniquely identify the item. Item type determines PairType, which can be useful in analyzing the data. There may be up to five item types. PairType is determined by the combination of the item types of the two items of the pair. If there is only one item type (as in fig. 2-1, where all item codes begin with “C”), there is only one pair type, and the variable PairType is uninformative (with only one item type, all pairs will be assigned a PairType of “1”). However, if there is more than one item type, there will be more than one pair type. For example, if there are three item types (say A, B, and C), there will be six pair types (AA, AB, AC, BB, BC, and CC). With six pair types, PairType codes will be the numerals 1 through 6. The PairType codes are listed in the Pairs.csv output file.

Following the item code is a field that allows the item to be excluded from being paired with selected other items when the set of paired comparisons is presented (see fig. 2-1). This field is particularly useful for avoiding presentation of pairs for which the choice is obvious, such as are comparisons of monetary amounts. The choice for pairs that are not presented to the respondent is determined by the program based on the order in which the items are listed in the fourth section of PAIRCOMP.TXT, with items further down in the list preferred to (chosen in favor of) items higher up in the list.

The next section of PAIRCOMP.TXT, ### DOUBLE SORT TIES BY, allows the user to select between two options for ordering items of identical preference score for input to the double-sort (DS) procedure. The DS procedure determines the inconsistent choices that may be presented for retrial. As described in Chapter 4, the DS procedure isolates inconsistent choices by re-ordering the rows and columns of the choice matrix based on preference score. A simple ordering by preference score has no effect on the order of items of identical preference score, but the order of items of identical preference score nevertheless affects the ability of the DS procedure to correctly isolate inconsistent choices. The DS procedure is most effective when items of identical preference score are arranged in order of increasing value on the choice dimension. Option A (which was used for the empirical studies described further in this chapter) is to rely on the order in which the items were listed in the input file (the fourth section of PAIRCOMP.TXT).

Option B (not available when the empirical studies were performed) is to make use of the finding, presented in Chapter 5, that later choices tend to be more consistent than earlier choices. This option weights the choices based on the order in which the choices were made, placing more weight on later choices, to resolve ties in preference score. Tied items of greater weighted score are placed after items of lower weighted score in the matrix submitted to the DS procedure. Regardless of which option (A or B) is selected, the program lists (in the Pairs.csv output file) the identified inconsistent choices for both options. If the user has a good idea of the order of the items on the choice dimension, option A may be best. Otherwise, option B is likely to be best.

The next section of PAIRCOMP.TXT, `### PRESENT RETRIALS?`, allows the user to forego or limit the presentation of retrials. Recall that the program can present for retrial all pairs determined by the DS procedure to be inconsistent plus 10 randomly selected consistent pairs. If retrials are to be presented, two parameters are entered. The first parameter is a threshold coefficient of internal reliability (ζ), below which retests will not be presented. An unusually inconsistent respondent may have a large number of inconsistent choices. This threshold level of ζ allows the retest process to be skipped for such a respondent. If the user wants retests to be presented for all respondents regardless of ζ , a very low threshold level, such as 0.01, is listed. If no inconsistent choices are found by the DS procedure, the 10 consistent pairs are not retested, regardless of the ζ listed.

The second parameter of the `### PRESENT RETRIALS?` section allows the user to limit the inconsistent pairs that are retested to those that meet a PSD threshold. This option is provided because some pairs identified by the DS procedure as inconsistent are likely to contain items of very similar value, which may not have a high probability of actually being inconsistent. For example, pairs identified as inconsistent that contain items of identical preference score ($\text{PSD} = 0$) have only about a 50 percent chance of actually being inconsistent. The user may wish to limit the inconsistent pairs that are retested to those with a greater likelihood of actually being inconsistent. A PSD threshold of 0 will allow all pairs identified by the DS procedure as inconsistent to be retested (subject to the respondent meeting the ζ threshold). A PSD threshold of 1 will exclude from retest all inconsistent pairs of $\text{PSD} = 0$, and so forth. Note that using a PSD threshold above 0 will limit the ability of retesting to correct inconsistent choices.

The final six sections of PAIRCOMP.TXT deal with the rating, demographic questions that may be presented, and the closing screen of the session. The instructions in figure 2-1 for these sections are self-explanatory.

Output files

As will be seen in the list of files that follows, several choice matrices are produced by the PC program. The “original choice matrix” contains the original (initial, not retest) respondent choices for the full set of pairs, with the items ordered in the matrix (both top to bottom and left to right) as they were listed in PAIRCOMP.TXT. Two resorted choice matrices are produced from the original choice matrix. The resorted matrices contain the same choices as the original choice matrix, but with the items reordered using the DS procedure. The two versions of the resorted choice matrix correspond to options A and B for ordering items of tied preference score before submitting the matrix to the DS procedure. The results of only one of these two resorted matrices are used to determine which choices will be retested (if retests were requested), but both resorted matrices are provided by the PC program (thus the user can see which choices would have been isolated as inconsistent if the other option had been selected). Corresponding to the three choice matrices just described are three other matrices that incorporate the retest choices. These matrices, of course, are produced only if retests were presented to the respondent. The “retested original choice matrix” is identical to the original choice matrix except that for each retested pair the retest choice replaces the original choice (note that the retest choice for a pair may be identical to the original choice). This retested original choice matrix is produced so that the user can see how the preference scores change when the retest choices are incorporated. From the retested original choice matrix, two resorted choice matrices are created, based again on options A and B for ordering items of identical preference score. These two final (retested sorted) matrices are produced so that the user can see how the set of inconsistent choices changes when the retest choices are used.²⁴

Each run of the program (thus, each respondent) produces a set of .csv files (.csv files easily import to Excel and other programs). The files, which are placed in a folder with a unique name beginning with the date of the run, are as follows:

- Pairs.csv — one line per pair of items, all followed by PairType codes. The order in which the pairs, one per line, are listed reflects the order of the items in the PAIRCOMP.TXT file. These lines have the following fields:
 1. Folder name where the .csv files are placed (the same for all pairs)
 2. Item code of row item
 3. Item code of column item
 4. Pair number (composed of the row number followed by the column number from the placement of the pair in the original choice matrix)
 5. Respondent choice (“1” if the column item was preferred to the row item, “0” otherwise)
 6. Time taken to make the original choice, in seconds
 7. Sequence (in other words, order of presentation) of all original pairs
 8. Procedure used for ordering items of tied preference score for the DS procedure for subsequent presentation of retest pairs (“1” if option A was used, “2” if option B was used) (blank if retests were not presented)

9. Retest (“0” if the pair was not retested, “1” if the pair was retested and was originally an inconsistent choice, “2” if the pair was retested and was originally a consistent choice) (blank if retests were not presented)
 10. Switched (“1” if the retested choice was different from the original choice, “0” otherwise) (blank if retests were not presented)
 11. Row number of the pair in the original choice matrix
 12. Column number of the pair in the original choice matrix
 13. Total of the row of the pair from the original choice matrix
 14. Total of the column of the pair from the original choice matrix
 15. Pair type (pair types are defined at the bottom of the Pairs.csv file, with one line per pair type listing the numeric code of the pair type along with the corresponding characters from the item code)
 16. Coefficient of internal reliability (the same for all pairs)
 17. Inconsistencies 1 (“1” if determined to be an inconsistent original choice by the DS procedure using option A for ordering items of tied preference score, “0” otherwise) (corresponds to SortedChoiceMatrix.csv)
 18. Inconsistencies 2 (“1” if determined to be an inconsistent original choice by the DS procedure using option B for ordering items of tied preference score, “0” otherwise) (corresponds to SequenceBasedSortedChoiceMatrix.csv)
 19. Inconsistencies 3 (“1” if determined to be an inconsistent choice by the DS procedure after retest choices replaced original choices in the choice matrix, with option A used for ordering items of tied preference score, “0” otherwise) (corresponds to RetestedSortedChoiceMatrix.csv) (blank if retests were not presented)
 20. Inconsistencies 4 (“1” if determined to be an inconsistent choice by the DS procedure after retest choices replaced original choices in the choice matrix, using option B for ordering items of tied preference score with retied pairs assigned the highest weights, “0” otherwise) (corresponds to RetestedSequenceBasedSortedChoiceMatrix.csv) (blank if retests were not presented)
- PreferenceScores.csv — one line per item listing the preference scores from OriginalChoiceMatrix.csv and RetestedOriginalChoiceMatrix.csv, with the items in the order they were listed in PAIRCOMP.TXT
 - RetestPairs.csv — lists the retest pairs in order of presentation (included only if retests were presented)
 - ItemRating.csv — ratings of the items and the sequence number of the rating question (included only if items were presented for ratings)
 - StatementRatings.csv — ratings of the statements and the statements that were presented (included only if statements were presented for ratings)
 - DemographicQuestionResponses.csv — answers to the demographic questions (included only if demographic questions were presented)
 - OriginalChoiceMatrix.csv — the raw choice matrix, as in figure 1-1a, with the items listed in the order they appear in PAIRCOMP.TXT
 - SortedChoiceMatrix.csv — the double-sorted choice matrix, as in figure 1-1b, with option A used to order items of tied preference score

- SequenceBasedSortedChoiceMatrix.csv — the double-sorted choice matrix, with option B used to order items of tied preference score
- RetestedOriginalChoiceMatrix.csv — similar to OriginalChoiceMatrix.csv, but retest choices replace original choices for all retested pairs (included only if retests were presented)
- RetestedSortedChoiceMatrix.csv — similar to SortedChoiceMatrix.csv, but retest choices replace original choices for all retested pairs (included only if retests were presented)
- RetestedSequenceBasedSortedChoiceMatrix.csv — similar to SequenceBasedSortedChoiceMatrix.csv, but retest choices replace original choices for all retested pairs (included only if retests were presented)
- OriginalResponseTimeMatrix.csv — matrix arranged as in the raw choice matrix, but each cell gives the time taken to make the choice, in seconds
- OriginalSequenceMatrix.csv — matrix arranged as in the raw choice matrix, but each cell gives the sequence number of the choice
- RetestedOriginalSequenceMatrix.csv — similar to OriginalSequenceMatrix.csv, but retested choices replace original choices for all retested pairs (included only if retests were presented)

PAIRCOMP6.EXE, PAIRCOMP.TXT as it appears in figure 2-1, DOLLAR AMOUNT TEMPLATE.PSD, and a set of jpg files that go along with the instructions in the PAIRCOMP.TXT of figure 2-1 are available at the following website: <http://www.fs.fed.us/rm/value/software.html>.

Paired Comparison Simulation (PCS) Program

Evaluation of the reliability of a set of paired comparison judgments as measured by the internal consistency of responses in an empirical paired comparison experiment is not possible because we do not know respondents' V s for the items in the choice set. For the same reason, we cannot identify the individual respondent choices that are not consistent with those V s. The commonly used measures of consistency are the number of circular triads (C) and the resulting coefficient of internal reliability (ζ). C and ζ are summary statistics, based on the full set of paired comparisons of a given respondent. As such, they do not indicate which choices are inconsistent (in other words, which choices cause circular triads).²⁵ In addition, a lack of circular triads does not necessarily imply true internal consistency. For example, an internally inconsistent choice between two items whose V s are ordinally adjacent will produce no circular triads.

The double-sort (DS) procedure attempts to address this lack of indication of which choices are inconsistent by using a respondent's preference scores as surrogates for the respondent's V s. As described fully in Chapter 4, the DS procedure reports as inconsistent those choices that are not consistent with the ordinal relationship among preference scores. Unfortunately, circular triads cause preference score ties, which introduce errors in the procedure.

Our primary objective in developing the PCS model was to test the DS procedure for isolating inconsistent choices. Only by simulating actual inconsistent choices and comparing them to the choices selected by the DS procedure as inconsistent were we able to assess the efficacy of the procedure.²⁶ The PCS model creates a set of virtual respondents that behave according to the law of comparative judgment, with true expected values (V) and normal discriminial dispersions, and make choices that are submitted to the PC program as if real people were performing the experiment. Because the V s are known, we can observe the relationship between true internally inconsistent choices (in other words, choices inconsistent with the V s) and the choices selected by the DS procedure as inconsistent (called “reported” inconsistent choices). Our test of the DS procedure investigates two relationships, those (1) between the number of circular triads and the number of truly internally inconsistent choices and (2) between the choices reported by the DS procedure as internally inconsistent and the choices that truly are internally inconsistent.

The PCS model also allowed us to explore several other issues, including the ability of different scaling procedures to produce an interval-scale metric of the underlying preferences, the effect of discriminial dispersion on the coefficient of internal reliability, the effect on scale values and on measures of reliability of the number of items in the choice set, and relationships among virtual respondents’ true expected values (V) and discriminial dispersions. In addition, as seen in later chapters, by comparing results of actual paired comparison experiments with results of simulation runs, we were able to obtain some clues about the relevance of the law of comparative judgment. Most importantly, we were able to gain insights about the variation among respondents in their discriminial dispersions and about change in discriminial dispersion over time as respondents become more familiar with the items being compared and perhaps also with their own preferences.

PCS Program Modules

The PCS model is made up of five key modules:

1. The interpersonal dispersion module generates the V s of each respondent for each item, allowing the V s for a given item to differ among respondents.
2. The instantaneous value module specifies the U s (see equation 3) for each item of each choice based on the V s of module 1 and the specified discriminial dispersions.
3. The within-pair difference module calculates the difference between the two U s of a given choice.
4. The consistency test module determines whether the choice is consistent with the V s.
5. The mistake module randomly generates mistakes based on the user’s specification of the probability of a mistake.

The runs of the simulation model performed for this monograph were based on homogeneous samples of respondents with respect to the V s, so we used the interpersonal dispersion module only to specify the V s, not to cause them to differ across respondents. In more detail, the other four modules are as follows:

Instantaneous value generator—The purpose of this module is to generate each respondent's normally distributed U s based on their V s and discriminial dispersions. Each respondent has a V for each item as generated in Module 1, but for each choice the respondent actually draws a value (U) for each item in the pair from a normal dispersion about the V for the item. This module uses a normally distributed random number generator to produce a pair of U_{ipc} each time a respondent makes a pair-wise choice, where $i = 1, 2, \dots, t = \text{number of items}$, $p = 1, 2, \dots, s = \text{number of virtual respondents}$, and c designates the choice being made (in other words, first choice, second choice, and so forth). The module saves a pair of U_{ipc} for each choice.

User input parameters—The user specifies Δ_i , the standard deviations of the discriminial dispersions of the various items. A normally distributed standard normal random number generator produces a standard normal deviate, x_{ipc} , for each respondent for each item in each paired choice. Another parameter, γ_p , described below, allows the discriminial dispersion to vary across respondents. $U_{ipc} = V_{ip} + x_{ipc}\Delta_i$ is the value used by respondent p for item i in choice c . It is generated anew every time an item i comes up in a paired comparison. Every choice requires two U_{ipc} .

γ_p Generator— γ_p is a random number that amplifies or diminishes Δ_i . In empirical experiments, some respondents are more reliable than others, implying that the variance of the discriminial dispersion varies across respondents. Empirical data imply that the variation of Δ_i across respondents is approximately log-normal. This submodule randomly generates a standard normal deviate, x_p , for each respondent. Then, $x_p = \text{Ln}(\gamma_p)$ and $\gamma_p = \text{Exp}(x_p)$. If the user does not want Δ to vary across respondents, γ is set to 1.0.

Within-pair difference calculator—This module calculates the $U_{ipc} - U_{ipc}$ in response to each pair-wise query presented by the PC program. This module saves the absolute value of the within-pair difference and the pair index for each choice. It also sends $U_{ipc} - U_{ipc}$ for each choice to Module 4.

Consistency test—In response to each pair-wise query presented by the PC program, this module chooses the item with the largest U_{ipc} in each pair and then compares the sign of $U_{ipc} - U_{ipc}$ with the sign of the corresponding $V_{ip} - V_{ip}$. If the signs differ (agree), the module flags the choice as inconsistent (consistent). The module saves the consistency designation and sends it and the choice to Module 5, the Mistake Generator. The inputs are the individual values generated by Module 2 and the individual respondent V s generated by Module 1.

Mistake generator—Data from previous real experiments suggest that in addition to making randomly inconsistent choices, a real respondent may also make mistakes. A mistake occurs when the respondent hits the wrong arrow key in response to a PC program query and fails to catch and correct the mistake. This module uses an independently and uniformly distributed random number generator to generate mistakes with $\pi(\text{mistake}) = w$. A mistake thus generated reverses the choice previously made by the decision-maker. This module saves whether each choice is a mistake or not and whether it is consistent or inconsistent, and then sends the choice to the PC program.

User input—The user specifies w . Empirical data suggest that w falls in the vicinity of 0.02.

A more detailed description of the PCS model is available from the first author.

Empirical Studies

The four studies that provide the real-respondent data for this paper each obtained judgments for a mixture of public goods, private goods, and amounts of money. Two of the studies used split-sample designs with different experimental materials for the different samples; thus the four studies contain a total of six experiments each with a set of paired comparison data from a unique group of respondents (table 2-1). Some goods were used in two or more experiments, but other goods were unique to a specific experiment. Across the six experiments, a total of 1,235 respondents, all university students, provided paired comparison judgments. Five respondents were later dropped for various reasons, reducing the total number of respondents to 1,230. All public goods were locally relevant and involved the university campus or its surrounding community. All private goods were common consumer items. In total, these six experiments yielded 170,546 usable original respondent choices, each between a pair of items, plus many retest choices.

Table 2-1. Key characteristics of the six real-people experiments.

Study	Subjects	Subjects dropped	Private goods	Public goods	Dollar amounts	Pairs / subject	Choices ^a
			number of				
Original	330	3	4	6	11	155	50,685
Clarke	464	1	4	5	9	117	54,171
Birjulin A	95	0	4	6	11	155	14,725
Birjulin B	95	1	4	6	11	155	14,570
Bid Range 700	125	0	5	5	10	145	18,125
Bid Range 9000	126	0	5	5	10	145	18,270
Total	1235	5					170,546

^a Total number of original choices (in other words, not including retests).

As an example of the methods of these studies, we summarize the methods used by Peterson and Brown (1998) in what we call the Original study. The choice set consisted of six public goods, four private goods, and 11 sums of money (21 items in total, table 2-2). Each respondent made 155 choices consisting of 45 choices between goods and 110 choices between goods and sums of money. They did not choose between sums of money (it was assumed that larger amounts of money were preferred to smaller ones), but did choose between all other possible pairs of the items.

Table 2-2. Items included in the Original study.

PRIVATE GOODS	
1.	A meal at a Fort Collins restaurant of your choice, not to exceed \$15. (Meal)
2.	Two tickets and transportation to <i>one</i> of the following: a. A Colorado ski area of your choice. b. A concert of your choice in Denver (contemporary or classical). c. A Broncos, Nuggets, or Rockies game. d. A cultural event of your choice at the Denver Center for the Performing Arts. Estimated value \$75. (Tickets)
3.	A nontransferable \$200 certificate for clothing at a Fort Collins store of your choice. (Clothes)
4.	A nontransferable certificate for you to make \$500 worth of flights on an airline of your choice. (Air Travel)
PUBLIC GOODS	
1.	A no-fee library service that provides videotapes of all course lectures so that students can watch tapes of lectures for classes they are not able to attend. (Videotape Service)
2.	Parking garages to increase parking capacity on campus such that students are able to find a parking place at any time, without waiting, within a 5-minute walk of any building at no increase in the existing parking permit fee. (Parking Capacity)
3.	CSU purchase of 2,000 acres in the mountains west of Fort Collins as a wildlife refuge for animals native to Colorado. (Parking Capacity)
4.	A CSU-sponsored on-campus springtime festival with a variety of live music and student participation events with no admission fee. (Spring Festival)
5.	Expansion of the eating area in the Lory Student Center to ensure that any student can find a seat at any time. (Eating Area)
6.	A cooperative arrangement between CSU, local business groups, and the citizens of the community that would ensure Fort Collins air and water would be at least as clean as the cleanest 1 percent of the communities in the United States. (Clean Arrangement)
MONEY AMOUNTS (\$)	
1, 25, 50, 75, 100, 200, 300, 400, 500, 600, and 700	

The four private goods of the Original study were familiar market goods with specified prices (table 2-2). For example, one of the goods was a meal at any local restaurant, not to exceed \$15. The six public goods were of mixed type. Two were pure public environmental goods — they were nonrival and nonexcludable in consumption. For example, one was a purchase by the University of 2,000 acres of land in the mountains west of town as a wildlife refuge for animals native to the area. The remaining four public goods were excludable by nature but stated as nonexcludable by policy. They were also nonrival until demand exceeds capacity. For example, one was a no-fee library service that provides video tapes of all course lectures. The two pure public goods benefited all people in the broader community, whereas the other four public goods benefited only the students. The 11 sums of money were \$1, \$25, \$50, \$75, and \$100 to \$700 in intervals of \$100. These amounts were derived from pilot studies in order to have good variation and distribution across the values of the goods.

Respondents were asked to choose one or the other item under the assumption that either would be provided at no cost to the respondent. The respondent simply chose the preferred item in each pair. If respondents were indifferent between the two items in a pair, they were still asked to make a choice. Indifference across respondents was later revealed as an equal number of choices of each item.²⁷

The experiment was administered on personal computers using PC.EXE, as described above in the “Paired Comparison (PC) Program” section. At the end of the 155 paired comparisons, the computer code repeated those pairs for which the individual’s choice was not consistent as determined by the DS procedure, plus 10 consistent choices.

The Clarke study (Clarke 1996, Clarke and others 1999) used identical or similar items to the Original study (table 2-3). The Clarke study used a split sample design with six groups of respondents. Groups differed in the priming information they were given to read prior to the paired comparisons, which varied across groups in amount of concern expressed for the environment and whether respondents were to assume that ultimate decisions would be based only on the respondent's individual responses or on a combination of all students' responses. The differences among the six treatments are not pertinent to our use of the data and are ignored herein.

Table 2-3. Items included in the Clarke study.

1. A meal at a Fort Collins restaurant of your choice, not to exceed \$15. (**Meal**)
2. Two tickets and transportation to *one* of the following:
 - a. A Colorado ski area of your choice.
 - b. A concert of your choice in Denver (contemporary or classical).
 - c. A Broncos, Nuggets, or Rockies game.
 - d. A cultural event of your choice at the Denver Center for the Performing Arts.Estimated value \$75. (**Tickets**)
3. A nontransferable \$200 certificate for clothing at a Fort Collins store of your choice. (**Clothes**)
4. A nontransferable certificate for you to make \$500 worth of flights on an airline of your choice. (**Air Travel**)

PUBLIC GOODS

1. An open, undeveloped area 1 mile wide from Fort Collins to Loveland that preserves wildlife habitat and maintains a natural greenbelt to separate areas of rapid urban expansion. (**Greenbelt Corridor**)
2. An interconnecting, well-maintained set of bicycle trails in Fort Collins that would enable safe and scenic access to campus from all parts of town. (**Bicycle trails**)
3. A program whereby all CSU trash is sorted to recover all recyclable materials before the trash is sent to the landfill, using paid CSU students to do the sorting. (**Recycling**)
4. CSU purchase of 2,000 acres in the mountains west of Fort Collins as a wildlife refuge for animals native to Colorado. (**Wildlife Refuge**)
7. A cooperative arrangement between CSU, local business groups, and the citizens of the community that would ensure Fort Collins air and water would be at least as clean as the cleanest 1 percent of the communities in the United States. (**Clean Air**)

MONEY AMOUNTS (\$)

1, 25, 50, 100, 300, 500, 700, 1000, 9000

Birjulin (1997) used a split sample design with somewhat different items in the two sets. His first set, called Birjulin A here, used identical or similar items to the Original study (table 2-4). For his other set (Birjulin B here), he used the same private goods as the Original study but with different dollar values and largely different public goods compared to the Original study (table 2-5). The differences between the two treatments are not pertinent to our use of the data and are ignored herein.

The Bid Range study data, used by Rosenberger and others (2003), included items either identical or very similar to the Original study (table 2-6). Respondents were separated into two groups that differed only in the set of monetary amounts included among the items, with one set using a much higher range (up to \$9,000) than the other (up to \$700) in order to examine the impact of bid range on the choices.

Table 2-4. Items included in the Birjulin A study.

PRIVATE GOODS	
1.	A meal at a Fort Collins restaurant of your choice, not to exceed \$15. (Meal/MEA)
2.	Two tickets and transportation to <i>one</i> of the following: a. A Colorado ski area of your choice. b. A concert of your choice in Denver (contemporary or classical). c. A Broncos, Nuggets, or Rockies game. d. A cultural event of your choice at the Denver Center for the Performing Arts. Estimated value \$75. (Tickets/TIC)
3.	A nontransferable \$200 certificate for clothing at a Fort Collins store of your choice. (Clothes/CLO)
4.	A nontransferable certificate for you to make \$500 worth of flights on an airline of your choice. (Air Travel/AIR)
PUBLIC GOODS	
1.	A no-fee library service that provides videotapes of all course lectures so that students can watch tapes of lectures for classes they are not able to attend. (Videotape Service/VID)
2.	Parking garages to increase parking capacity on campus such that students are able to find a parking place at any time, without waiting, within a 5-minute walk of any building at no increase in the existing parking permit fee. (Parking Capacity/PRK)
3.	CSU purchase of 2,000 acres in the mountains west of Fort Collins as a wildlife refuge for animals native to Colorado. (Wildlife Refuge/WLD)
4.	A CSU-sponsored, on campus springtime weekend festival with a variety of live music and student participation events with no admission fee. (Spring Festival/SPR)
5.	Expansion of the eating area in the Lory Student Center to ensure that any student can find a seat at any time. (Eating Area/EAT)
6.	A cooperative arrangement between CSU, local business groups, and the citizens of the community that would ensure Fort Collins air and water would be at least as clean as the cleanest 1 percent of the communities in the United States. (Clean Arrangement/CLE)
MONEY AMOUNTS (\$)	
1, 25, 50, 75, 100, 200, 300, 400, 500, 600, and 700	

Table 2-5. Items included in the Birjulin B study.

PRIVATE GOODS	
1.	A meal at a Fort Collins restaurant of your choice, not to exceed \$40. (Meal/MEA)
2.	Two tickets and transportation to <i>one</i> of the following: a. A Colorado ski area of your choice. b. A concert of your choice in Denver (contemporary or classical). c. A Broncos, Nuggets, or Rockies game. d. A cultural event of your choice at the Denver Center for the Performing Arts. Estimated value \$100. (Tickets/TIC)
3.	A nontransferable \$400 certificate for clothing at a Fort Collins store of your choice. (Clothes/CLO)
4.	A nontransferable certificate for you to make \$1000 worth of flights on an airline of your choice. (Air Travel/AIR)
PUBLIC GOODS	
1.	Improved landscaping, including art, artistic architecture, and areas where students could relax or study outside, in the main courtyard between the Lory Student Center, Morgan Library, Clark building, Glover building, and the engineering building. (Landscaping/LSP)
2.	Parking garages to increase parking capacity on campus such that students are able to find a parking place at any time, without waiting, within a 5-minute walk of any building at no increase in the existing parking permit fee. (Parking Capacity/PRK)
3.	A commitment to make CSU the leading undergraduate institution in the Rocky Mountain region, such that a degree from CSU would assure you the employment of your choice. (Leading Institution/INS)
4.	Offering a greater number of sections of the same course each semester, at a greater variety of times so that students could get into the courses they need at times that are convenient to them. (Courses/CSU)
5.	An on-campus bicycle path system to all buildings that would separate bicycle traffic from pedestrian and automobile traffic, including more bicycle racks outside each building. (Bike Paths/BIK)
6.	A cooperative arrangement between CSU, local business groups, and the citizens of the community that would ensure Fort Collins air and water would be at least as clean as the cleanest 1 percent of the communities in the United States. (Clean Arrangement/CLE)
MONEY AMOUNTS (\$)	
1, 25, 50, 75, 100, 200, 300, 400, 500, 600, and 700	

Table 2-6. Goods included in the Bid Range study.

PRIVATE GOODS	
1.	A meal at a Fort Collins fast food or cafeteria-style restaurant of your choice.
2.	A \$25 gift certificate for use at a bookstore of your choice.
3.	Two tickets to ONE of the following: a. A concert of your choice in Denver (contemporary, new age, hard rock, or classical). b. General admission to a professional sport event in Denver (Broncos, Rockies, Nuggets, or Avalanche).
4.	A nontransferable \$200 certificate for clothing at a Fort Collins store of your choice.
5.	A nontransferable certificate for a round-trip flight to any major city in the contiguous 48 states on an airline of your choice.
PUBLIC GOODS	
1.	A no-fee library service that provides videotapes of all course lectures so that students can watch tapes of lectures for classes they are not able to attend.
2.	Parking garages to increase parking capacity on campus such that students are able to find a parking place at any time, without waiting, within a 5-minute walk of any building at no increase in the existing parking permit fee.
3.	CSU purchase of 2,000 acres in the mountains west of Fort Collins as a wildlife refuge for animals native to Colorado.
4.	Expansion of the eating area in the Lory Student Center to ensure that any student can find a seat at any time.
5.	A cooperative arrangement between CSU, local business groups, and the citizens of the community that would ensure Fort Collins air and water would be at least as clean as the cleanest 1 percent of the communities in the United States.
MONEY AMOUNTS (\$)	
Treatment 1: 10, 30, 50, 70, 100, 200, 300, 400, 500, and 700	
Treatment 2: 10, 30, 50, 70, 100, 300, 700, 1000, 3000, and 9000	

Item input order versus aggregate item rank order. As described in Chapter 4, the order of items in the raw choice matrix makes a difference in how well the DS procedure identifies inconsistent choices between items with identical preference scores. The order of the items in the raw choice matrix was set by the order in which the items were listed in the PAIRCOMP.TXT file. The DS procedure more accurately identifies inconsistent choices when the items of identical preference score are arranged in the raw choice matrix in order of increasing value on the dimension of interest. We were not aware of this issue when we designed the original PC program and administered the real-people experiments reported herein. Thus the input order of the items was generally not the most effective order for the DS procedure.

In all six of the real-people experiments, at least half of the items were dollar magnitudes. Respondents did not choose between dollar magnitudes, but only between goods or between a good and a dollar amount. It was assumed that respondents always would choose larger dollar magnitudes over smaller dollar magnitudes.

Dollar magnitudes followed the goods in input order in each study and the input order of the dollar magnitudes was always ascending.²⁸ With the exception of the Clarke study, the relationship between input order of the goods and final aggregate rank order was approximately random. Goods in the Clarke study tended to be entered in order of increasing value. Thus it appears that the input order of items in each of the real-people studies was effectively random or modestly ascending because of the effect of the dollar magnitudes, which were always in ascending order.

Statistical Procedures

Unless stated otherwise, when we report an R^2 it is an adjusted R^2 .

Summary and Conclusions

We have described two computer programs, one for administering paired comparisons to real respondents and the other for simulating the paired comparison responses of virtual respondents based on the assumption that they act according to the Law of Comparative Judgment (most importantly that their discriminial dispersions are normally distributed). An early version of the former program was used in the six experiments with real respondents, which provided over 170,000 actual choices. Of course, there is no limit to the number of virtual respondents whose choices may be simulated, and a great many were used, as seen in later chapters. Each set of data—the real choices and the simulated choices—allow exploration of important issues, but perhaps the most interesting issue is whether actual choices conform to the law of comparative judgment (which is examined by the comparison of the real and simulated choices). Such a comparison allows a look at the extent to which real respondents act according to the theoretical model upon which the study of paired comparisons is based. It is this question of behavioral compliance with the underlying distributional assumptions that has often been set aside—both in psychology and in applications of economic random utility theory, whether in choice prediction models or discrete choice contingent valuation—in the pursuit of estimates of scale values.

Chapter 3. Scaling

In this chapter, we use the paired comparison simulation (PCS) model to investigate the ability of several scaling procedures to discover the expected values (V) of the items being judged. This is accomplished by simulating the choices of a large set of virtual respondents systematically placed under alternative conditions. Recall from Chapter 2 that with the simulation model—unlike in a real-world experiment—we know the mean (V) and standard deviation (Δ) of the discriminial dispersions for each item judged by each respondent so that we can observe exactly how well the scaling procedures perform assuming people act according to the assumptions of the model (most importantly, that the dispersions are all normally distributed).

These simulations explore two main issues. First, the simulations examine the effect of the placement of the V s along the preference continuum on the ability of the scaling procedures to detect the relative placement of the V s. This is approached by examining three arrangements of the V s: evenly spaced, randomly spaced, and clustered into distinct groups. Second, the simulations examine the effect of relaxing the assumption of identical discriminial dispersions for the items. This is accomplished by systematically varying Δ from 1 to 10 within the context of V s ranging from 1 to 21 on the preference continuum. Further simulations address these two issues together. In most of the simulations, all respondents have identical Δ s, but for some, the respondents are organized in groups, each with a different set of Δ s. For most simulations, a very large sample of virtual respondents is used in order to investigate the expected result when sample size is not a constraint. At the end of the chapter we introduced the sample size constraint.

Scaling Procedures

The scaling procedures investigated are mean preference scores (MPSs) and two psychometric scaling procedures — Thurstone's Case V scale and a version of a logistic transformation scale. Computation of each of these scales begins with the t -by- t frequency matrix (see fig. 1-4), wherein each cell, f_{rc} , gives the frequency across the s respondents in the sample, with which the item in column c was chosen above the item in row r . Then the MPS of the item in column c (MPS_c) is computed as:²⁹

$$MPS_c = \sum_{r=1}^t f_{rc} / s \quad (16)$$

Thurstone's scaling model is described generally in Chapter 1 and in detail in other sources (for example, Bock and Jones 1968, Torgerson 1958). Assuming Case V, computation of scale values is straightforward. Once the frequency matrix is available, it requires only three steps. First, each frequency is converted to a proportion (ρ):

$$\rho_{rc} = f_{rc} / s \quad (17)$$

where ρ_{rc} is the proportion of respondents who choose the column item over the row item. Next, each proportion is converted to a standard normal deviate:

$$x_{rc} = \Phi(\rho_{rc}) \quad (18)$$

Finally, the mean of the standard normal deviates in a given column (T_c) yields the scale value for the column item:

$$T_c = \sum_{r=1}^t x_{rc} / t \quad (19)$$

Importantly, Thurstone's model assumes that the disturbance distributions of all items overlap to some degree, an assumption that is formalized in the adoption of normally distributed disturbance distributions (the normal curve being asymptotic to the horizontal axis). Because the relation of unit normal deviates to their respective probabilities is approximately linear except at the tails of the normal curve, the effect of the transformation from proportions to unit normal deviates is of consequence only when the proportion matrix contains cells with relatively high (above about 0.9) or low (below about 0.1) proportions. The conversion to unit normal deviates essentially weights the lowest and highest proportions more heavily than it does less extreme proportions. Matrices with few very high or very low proportions produce scale values that correlate very highly with the MPSs, which are themselves a linear function of the proportions. For the Thurstone Case V approach to make a unique contribution, over and above MPSs, accurate estimation of the proportions involving the most and least preferred items is important.

Even if the assumption of completely overlapping disturbance distributions is met, however, it may not be evident in the data. Empirically, lack of overlap is indicated by the number of cells in the frequency matrix with a minimum (in other words, 0) or maximum (in other words, equal to the number of respondents, s) frequency, with corresponding proportions of 0 and 1.0, respectively. Cells with these values (called vacant cells) are less likely to occur as the sample of respondents gets larger, the variability in preference among those respondents gets greater, and the respondents' discriminial dispersions get larger.³⁰ When vacant cells occur, their unit normal deviates must be approximated, as vacant cells in the choice matrix produce infinite standard normal deviates (Bock and Jones 1968; Guilford 1954; Maxwell 1974; Torgerson 1958), but any approximation is to some extent arbitrary and no approximation is totally satisfying. To fill vacant cells in the frequency matrix, we chose to substitute "1" for "0" and $s - 1$ for s . Thus, for example, with a sample of 2,000 respondents, frequencies of 0 and 2,000 would be replaced with 1 and 1,999, respectively, yielding corresponding proportions in the proportion matrix of 0.0005 and 0.9995. We will not delve deeply into the vacant cell issue here, but suffice to say that, in general, the greater the number of vacant cells, the less the degree to which the resultant scale values reveal the true distances between the items (in other words, the less the scale values can be accepted as an accurate interval scale measure).

The other psychometric scaling procedure has been known as the BTL model, after three early proponents of the approach (Bradley and Terry 1952, Luce 1959).

This approach uses a logistic, as opposed to a normal, transformation to generate scale values from the choice matrix, and like Thurstone's Case V relies on the assumption of equal discriminial dispersions across stimuli. Called the Maxwell logistic model herein (for the source where we originally encountered the method), the approach converts all cells of the frequency matrix to proportions and then computes the logit (z) of the choice proportion for each pair of items as follows:

$$z_{rc} = \log_e (\rho_{rc} / (1 - \rho_{rc})) \quad (20)$$

where ρ_{rc} is as defined in equation 17. Alternatively, if vacant cells are present in the aggregate choice matrix, the Maxwell approach uses the following substitution:

$$z_{rc} = \log_e ((s_{rc} + \frac{1}{2}) / (s - s_{rc} + \frac{1}{2})) \quad (21)$$

where s_{rc} is the number of respondents who choose the item in column c over the item in row r (Maxwell 1974). Finally, the mean of the z_{rc} in a given column (M_c) yields the scale value for the column item:

$$M_c = \sum_{r=1}^t z_{rc} / t \quad (22)$$

This model has been found to give very similar results to the Thurstone (normal distribution-based) Case V model (Bock and Jones 1968, Davidson and others 1995), which is not surprising in that they are essentially equivalent except that the logistic density function of the Maxwell model replaces the Gaussian density function of Thurstone's approach. We will present our own comparison of the two approaches in this chapter.

In the next two sections we look at the effect of alternative arrangements of the V s, first assuming IID discriminial dispersions and then allowing for heterogeneous dispersions, on the estimation of scale values.

IID Discriminal Dispersions

Evenly Spaced V s

We first investigate the relation of V to scale value for identically and independently distributed (IID) normal discriminial dispersions with evenly spaced V s. Figure 3-1 shows the relation of V to MPS for a range of discriminial dispersion standard deviations (Δ). The V s of the 21 items are equal to the integers from 1 to 21, and all respondents have identical V s for a given item. Each of the 10 lines in figure 3-1 shows results for a different simulation involving 6,000 virtual respondents (or cases). Each line represents a different value for Δ , which is varied from 1 to 10, and each dot along a line indicates the value of an item.

The curves of figure 3-1 rotate counter-clockwise as Δ increases. The counter-clockwise rotation occurs because, as Δ increases, overlap among the discriminial dispersions of different items increases, so that items away from the center point of the preference continuum (in other words, away from $V = 11$ in fig. 3-1) are less likely to be judged in any given comparison as preferred (at the high end) or not preferred (at the low end). Each curve is approximately linear.

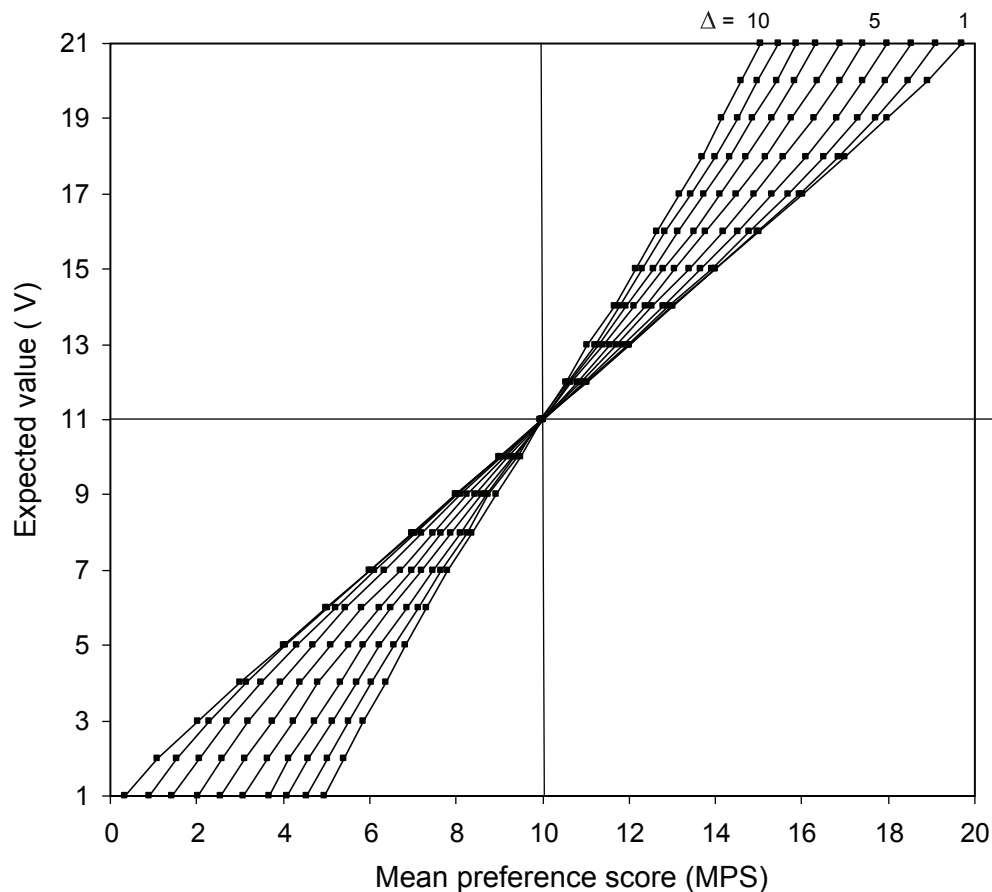


Figure 3-1. Relation of V to MPS as a function of Δ , 21 items (dots show data points for the items).

Other simulations showed that the general relationship (rotation and nearly straight lines) does not depend on the number of items in the choice set.

Table 3-1 lists the correlations of the scale values with the V s across the 21 evenly spaced items, revealing that the R^2 s exceed 0.99 for the three scaling procedures (MPS, T , and M) regardless of Δ . Clearly, given the conditions of these simulations—most importantly, many respondents and numerous items with evenly spaced V s—all three scaling procedures all produce excellent representations of the V s.

Table 3-1. R^2 for scale values versus V s, 21 items, 6,000 cases.

Scale	Δ		
	1	5	10
Evenly-spaced V s			
MPS	.9998	.9980	.9984
T	.9983	.9999	.9997
M	.9973	.9993	.9999
Randomly-spaced V s			
MPS	.9554	.9904	.9988
T	.9664	.9999	.9999
M	.9701	.9986	.9998

As shown in table 3-2, when $\Delta = 1$ there are many vacant cells (cells where one item in the pair was never chosen) in the frequency matrix, even though the matrix is based on 6,000 cases. At $\Delta = 5$ (table 3-3), and thus at $\Delta > 5$, there are no longer any vacant cells in the frequency matrix given 6,000 respondents with identical V s. However, as the R^2 s in table 3-1 indicate, even when vacant cells dominate the frequency matrix (as at $\Delta = 1$), the adjustments we use for vacant cells do not interfere with the ability of the psychometric scales (T and M) to accurately reflect the interval distances between the V s when the items are evenly spaced.

We conclude, therefore, that if the V s of the items in a paired comparison experiment are evenly spaced, discriminial dispersions are IID normal, and the sample of respondents is large, the MPSs comprise a nearly perfect interval scale of the V s, and under these ideal circumstances the psychometric scales offer no practical advantage over MPSs.

Table 3-2. Aggregate choice matrix, evenly spaced V s, $\Delta = 1$, 6,000 cases.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1																					
2	1467																				
3	486																				
4	111	458	1427																		
5	13	130	457	1443																	
6	0	10	107	460	1411																
7	0	0	15	112	483	1411															
8	0	0	0	13	91	475	1431														
9	0	0	0	0	19	101	454	1436													
10	0	0	0	0	1	12	88	460	1438												
11	0	0	0	0	0	1	15	95	490	1449											
12	0	0	0	0	0	0	3	16	113	512	1410										
13	0	0	0	0	0	0	0	0	7	103	468	1417									
14	0	0	0	0	0	0	0	0	1	13	91	465	1437								
15	0	0	0	0	0	0	0	0	0	2	13	114	462	1428							
16	0	0	0	0	0	0	0	0	0	0	2	13	92	429	1468						
17	0	0	0	0	0	0	0	0	0	0	0	2	11	105	437	1379					
18	0	0	0	0	0	0	0	0	0	0	0	0	0	19	112	463	1390				
19	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	108	469	1461			
20	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	18	101	448	1419		
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	98	427	1421	
	2077	6603	12048	18032	23962	30012	35970	41997	48039	54080	59934	65957	72007	77976	84007	89965	96037	102023	107799	113433	118042
	0.3462	1.1005	2.008	3.0053	3.9937	5.002	5.995	6.9995	8.0065	9.0133	9.989	10.993	12.001	12.996	14.001	14.994	16.006	17.004	17.967	18.9055	19.6737

Key:

true expected value
item number

item number

choice matrix (number of times
the column item was preferred
to the row item)

aggregate preference score
mean preference score

Table 3-3. Aggregate choice matrix, evenly spaced Vs, $\Delta = 5, 6,000$ cases.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
2	3344	3657	3958	4290	4551	4828	5049	5236	5400	5523	5634	5692	5799	5870	5908	5934	5943	5969	5974	5993	
3	2656		3384	3654	3912	4198	4537	4760	5046	5243	5419	5518	5626	5750	5801	5863	5910	5934	5960	5970	5975
4	2343	2616		3303	3573	4064	4274	4506	4747	5050	5207	5400	5526	5634	5741	5801	5865	5908	5925	5947	5969
5	2042	2346	2697		3288	3634	4011	4341	4590	4756	5069	5227	5411	5533	5646	5749	5773	5846	5887	5934	5953
6	1710	2088	2427	2712		3385	3743	4054	4260	4582	4785	5044	5241	5346	5520	5613	5738	5760	5848	5880	5917
7	1449	1802	1936	2366	2615		3389	3674	3984	4291	4555	4853	5048	5245	5382	5508	5629	5699	5804	5864	5896
8	1172	1463	1720	1989	2257	2611		3337	3657	3937	4276	4531	4802	5022	5204	5417	5505	5627	5754	5806	5864
9	951	1240	1494	1659	1946	2326	2663		3299	3610	3925	4239	4522	4855	5036	5249	5398	5526	5622	5740	5795
10	764	954	1253	1410	1740	2016	2343	2701		3255	3658	3998	4302	4528	4822	5058	5216	5414	5526	5634	5705
11	600	757	950	1244	1418	1709	2063	2390	2745		3339	3705	3983	4294	4602	4837	5034	5234	5393	5528	5640
12	477	581	793	931	1215	1445	1724	2075	2342	2661		3318	3685	3962	4264	4588	4820	4998	5230	5413	5521
13	366	482	600	773	956	1147	1469	1761	2002	2295	2682		3297	3622	4015	4295	4557	4801	5032	5201	5425
14	308	374	474	589	759	952	1198	1478	1698	2017	2315	2703		3345	3659	3955	4276	4632	4834	4977	5225
15	201	250	366	467	654	755	978	1145	1472	1706	2038	2378	2655		3376	3608	3965	4226	4580	4837	5039
16	130	199	259	354	480	618	796	964	1178	1398	1736	1985	2341	2624		3324	3633	4020	4296	4562	4782
17	92	137	199	251	387	492	583	751	942	1163	1412	1705	2045	2392	2676		3282	3609	4007	4289	4573
18	66	90	135	227	262	371	495	602	784	966	1180	1443	1724	2035	2367	2718		3368	3684	3976	4300
19	57	66	92	154	240	301	373	474	586	766	1002	1199	1368	1774	1980	2391	2632		3336	3668	3934
20	31	40	75	113	152	196	246	378	474	607	770	968	1166	1420	1704	1993	2310	2664		3410	3618
21	26	30	53	66	120	136	194	260	366	472	587	799	1023	1163	1438	1710	2024	2332	2590		3376
21	7	25	31	47	83	104	136	205	295	360	479	575	775	961	1218	1427	1700	2066	2382	2624	
	15448	18884	22601	26267	30347	35011	40043	44905	49703	54535	59957	65222	70232	75304	80321	85013	89207	93607	97659	101234	104500
	2.5747	3.1473	3.7668	4.3778	5.0578	5.8352	6.6738	7.4842	8.2838	9.0892	9.9928	10.87	11.705	12.551	13.387	14.169	14.868	15.601	16.277	16.8723	17.4167

Key:

true expected value
item number

choice matrix (number of times
the column item was preferred
to the row item)

aggregate preference score
mean preference score

Randomly Spaced Vs

Figure 3-2 shows the relation of V to MPS for a set of 21 items with randomly spaced Vs ranging from 1 to 21. For figure 3-2 we chose, from among several sets of random spacing we generated, one with relatively large differences from even spacing. As with the evenly spaced Vs of figure 3-1, the simulations are each based on choices of 6,000 virtual respondents, with each simulation representing a different Δ ranging from 1 to 10.

Again, the relationship rotates counter-clockwise as Δ increases, reflecting the greater overlap of the discriminial dispersions as Δ increases (fig. 3-2). With this particular configuration of randomly spaced Vs, eight of the items are bunched at the low end of the V range (all with $V < 4.5$). This causes significant overlap of the discriminial dispersions at even low levels of Δ , thus shifting the MPSs of those least-preferred items to the right as compared with the evenly spaced Vs of figure 3-1. Conversely, the Vs of the most preferred items are more widely spaced as compared with the evenly spaced Vs, and the MPSs for these items tend to be shifted to the right compared with evenly spaced Vs (for example, at $\Delta = 10$ the MPS of the most preferred item is about 15 for evenly spaced Vs and 16 for the randomly spaced Vs). With this randomly spaced set of Vs, the median V is only 7.8, not 11 as in figure 3-1. This downward shift in the median stretches out the curves for the more preferred items and leads to non-linearity of the curve in that upper region. The tendency toward non-linearity declines as Δ increases.

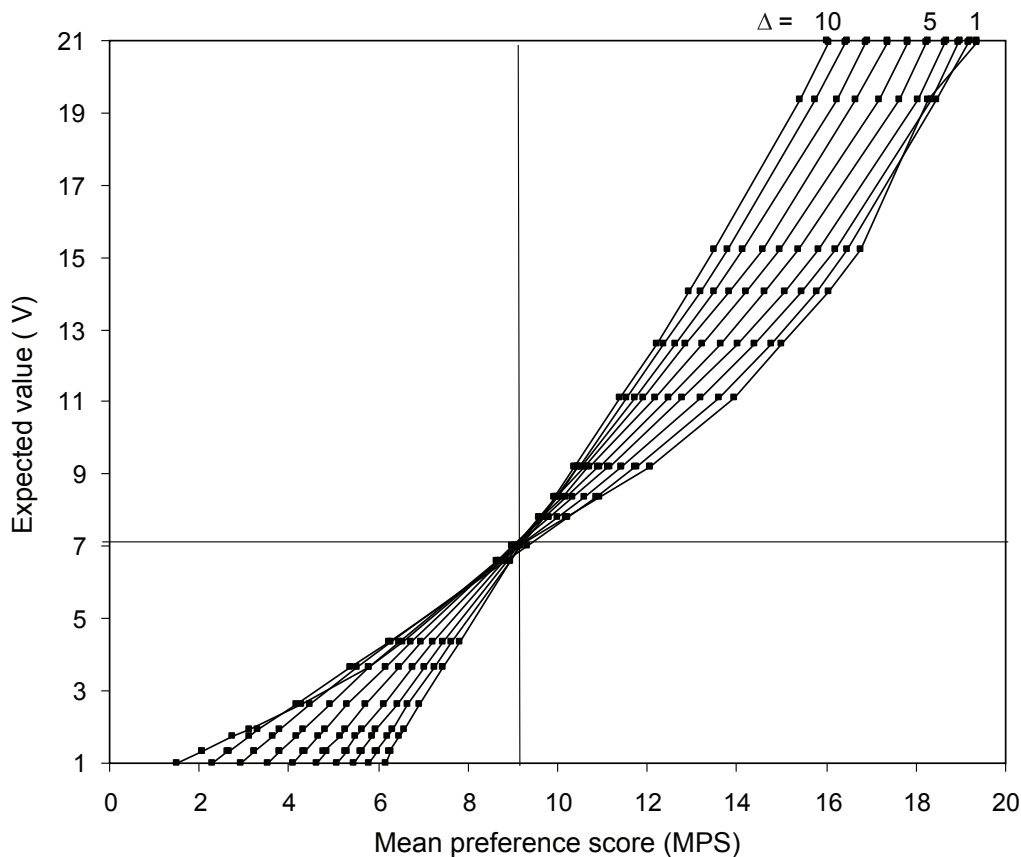


Figure 3-2. Randomly spaced Vs, 21 items, 6,000 cases per simulation (dots show data points for the items).

Figure 3-3 compares the three sets of scale values with the V s for the 21 items, with the scale values standardized to range from 0 to 1. If the scaling procedures performed perfectly, their values would fall on the straight line. As seen, the scaling procedures perform better at $\Delta = 5$ (fig. 3-3b) than at $\Delta = 1$ (fig. 3-3a). Figure 3-3 demonstrates at least three things of interest. First, for every item, the two psychometric scale values are quite similar. Second, the MPSs perform more poorly than the psychometric scales for almost all items for both levels of Δ , although the difference is small for most items. Third, the accuracy of the scale values improves with Δ . At $\Delta = 5$, the psychometric procedures very accurately reflect the V s.

Table 3-1 lists the correlations of the scale values with the V s for three levels of Δ (1, 5, and 10) for the particular set of randomly spaced V s depicted in figure 3-2. At $\Delta = 1$, the R^2 s are in the vicinity of 0.96, but by $\Delta = 5$, the R^2 s are all above 0.99. Whether the degree of non-linearity at low Δ is important depends, of course, on the demands of the application.³¹ In this context, it should be noted that error introduced by non-linearity when the V s are randomly spaced may be less important than other sources of error, such as heterogeneity of the sample.

As with evenly spaced V s (table 3-2), the frequency matrix for randomly spaced V s contains many vacant cells when $\Delta = 1$. These vacant cells again require arbitrary adjustment before the T and M scales can be estimated. In the case of the randomly spaced V s of figure 3-2, the psychometric scale values derived from the adjusted frequency matrix retain most of the curvature depicted for the MPSs.

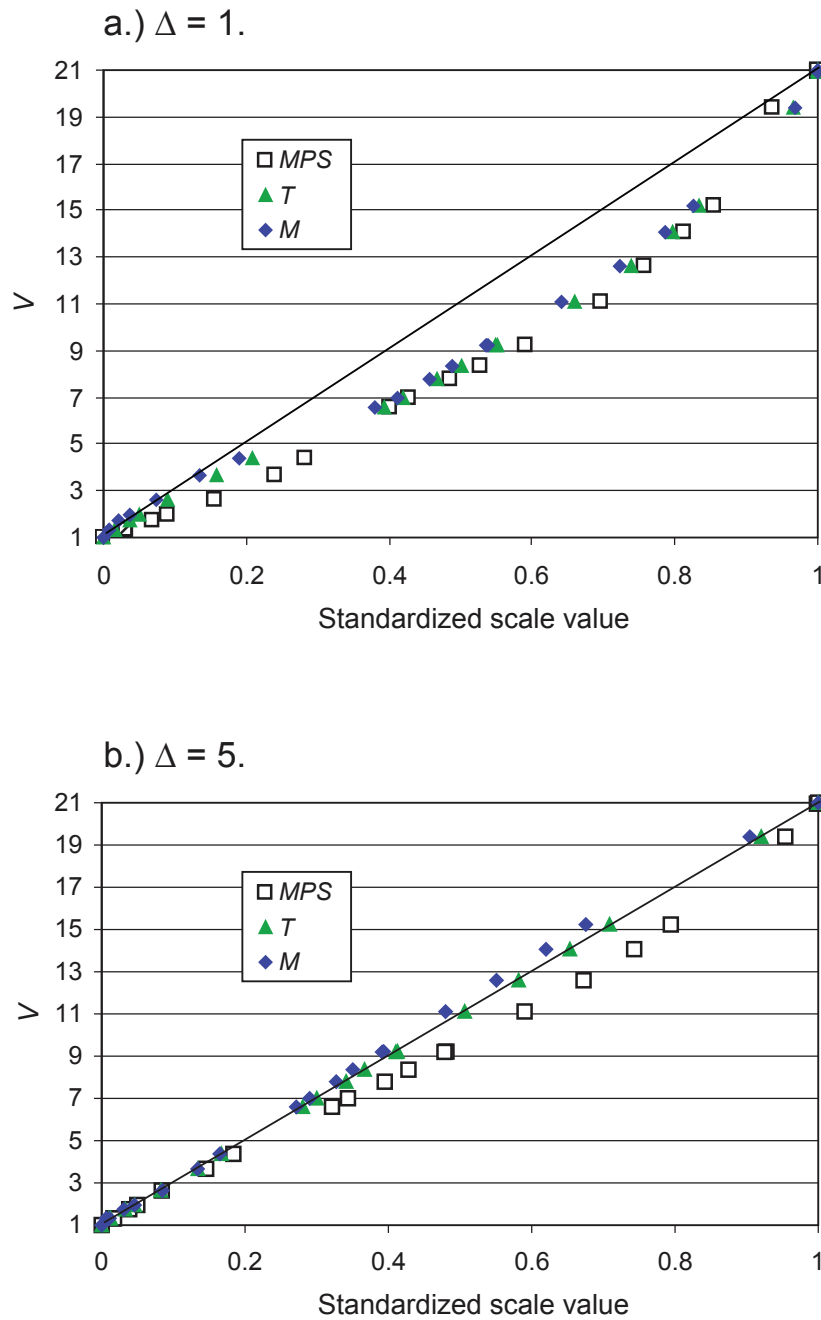


Figure 3-3. Comparison of scale values with V for 21 randomly spaced V s.

We surmise that the arbitrary adjustment of the choice matrix to fill in the vacant cells prevents the scaling methods from producing fully accurate interval scales at $\Delta = 1$. Moving to a Δ of 5, the increased dispersion has reduced the curvature of the MPS scale, and both the T and M procedures produce nearly perfect interval scales. The improvement over MPSs in linearity gained by the scaling methods is small, however (table 3-1).

We conclude that when V s are randomly spaced, the problems caused by vacant cells tend to limit the effectiveness of the psychometric scaling procedures just when the procedures would otherwise offer the most improvement (that is, when MPSs fail to provide an accurate measure). When the discriminial dispersion is sufficiently large relative to the range of V s such that vacant cells are not a problem, the MPSs come very close to comprising an interval scale. When the dispersion is small relative to the V range, such that empty cells are a problem, the need for arbitrary adjustment of the aggregate choice matrix to fill the vacant cells prevents the psychometric scaling methods from providing a substantial improvement over MPSs.³²

Vs in three clusters

A well-distributed but essentially random spacing of V s, similar to our random spacing simulation, is probably most likely in an empirical experiment, but a bunched or clustered arrangement is always a possibility. Because clustering has implications for the ability of scaling procedures to represent the underlying V s, we present several clustering possibilities (in this and later sections). In these cluster simulations, we again use 21 items with V s ranging from 1 to 21.

First we simulate a situation with items in three evenly spaced clusters centered at V s of 2, 11, and 20, with the seven items within a cluster spaced evenly, $0.2V$ apart. We performed separate simulations for Δ at 1, 2, 6, and 11, with each simulation using 2000 virtual respondents.³³

At all levels of Δ , the MPSs and the psychometric scales all comprise a nearly perfect interval scale representation of the V s ($R^2 > 0.99$, table 3-4). However, for MPSs, the overall R^2 masks an interesting anomaly: the slopes of the three within-cluster scales are quite different from the overall slope. The overall slope, that across all 21 items, is about 1.0. At $\Delta = 1$, the slopes of the three within-cluster scales are about 0.44 (fig. 3-4a). At this low level of Δ , there is no interaction between clusters (table 3-5). For example, an item in the lowest cluster is never preferred to an item in the middle cluster (given that the dispersions are normally distributed, some interaction was possible—

Table 3-4. R^2 for scale values versus V s, homogeneous discriminial dispersions, three clusters.

Scale	Δ			
	1	2	6	11
Clusters evenly spaced, V s evenly spaced within clusters				
MPS	.9937	.9996	.9999	.9999
T	.9999	.9998	.9999	.9999
M	.9996	.9993	.9998	.9999
Clusters randomly spaced, V s randomly spaced within clusters				
MPS	.9061	.9413	.9987	.9999
T	.9325	.9981	.9999	.9999
M	.9523	.9978	.9998	.9999

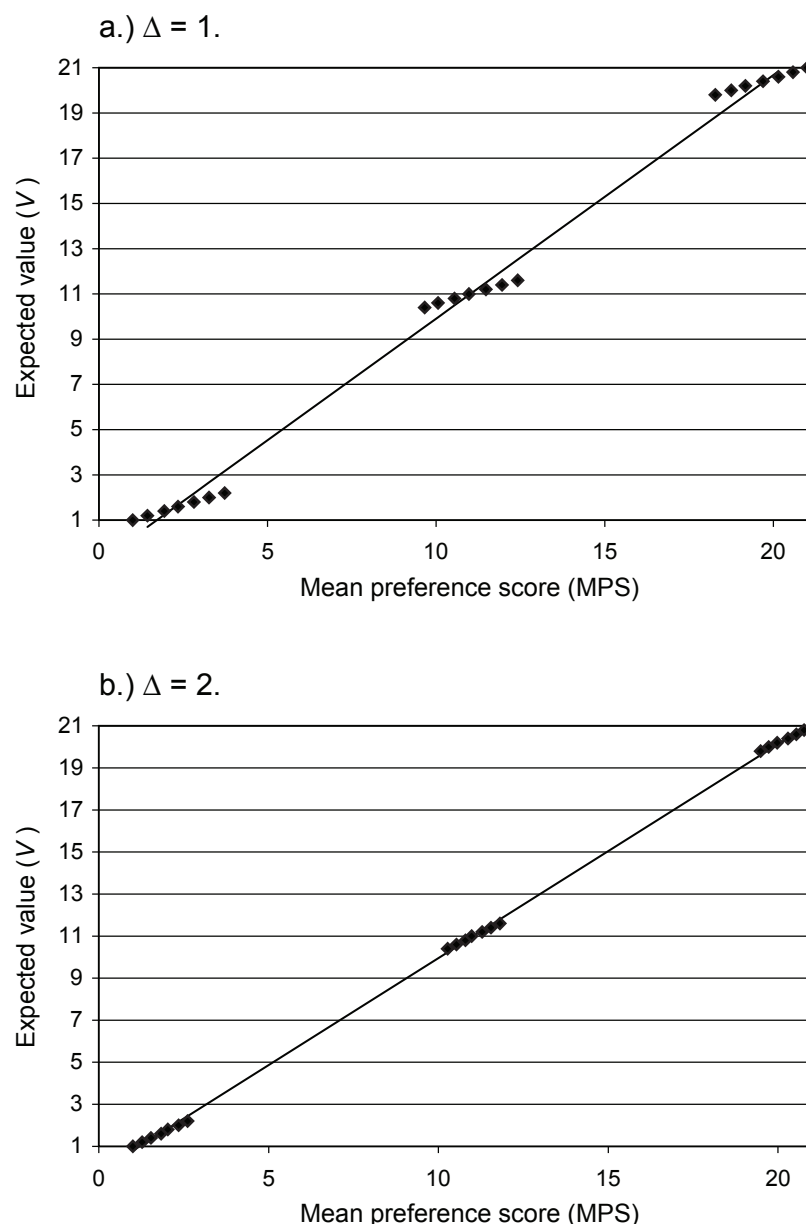


Figure 3-4. Relation of V to MPS, three evenly spaced clusters and evenly spaced V s within clusters, 21 items.

although very unlikely—but none occurred in this simulation). Each cluster is analyzed separately, as if the other clusters did not exist, to determine MPS of the items within a cluster. Thus, the within-cluster slope is not related to the overall slope. At larger levels of Δ , some interaction between clusters occurs and the within-cluster slopes become more similar to the overall slope. At $\Delta = 2$, the slopes of the three within-cluster scales are about 0.75 (fig. 3-4b), and by $\Delta = 6$, the within-cluster slopes are all about 1.0.³⁴

The T and M scales largely remove the differences between the within-cluster slopes and the overall slope, aligning the within-cluster slopes almost perfectly along the overall slope, thus offering an improvement over MPSs. This occurred despite the fact that at low Δ s, it was necessary to adjust the choice matrix for many vacant cells (table 3-5).

Table 3-5. Aggregate choice matrix, Vs in three evenly spaced clusters, evenly spaced within clusters, $\Delta = 1, 2,000$ cases.

	1.0	1.2	1.4	1.6	1.8	2.6	2.2	10.4	10.6	10.8	11.0	11.2	11.4	11.6	19.8	20.0	20.2	20.4	20.6	20.8	21.0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1		1122	1224	1310	1454	1509	1581	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
2	878		1149	1214	1303	1403	1546	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
3	776	851		1131	1178	1329	1423	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
4	690	786	869		1125	1233	1324	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
5	546	697	822	875		1081	1245	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
6	491	597	671	767	919		1092	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
7	419	454	577	676	755	908		2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
8	0	0	0	0	0	0	0		1091	1179	1339	1421	1546	1614	2000	2000	2000	2000	2000	2000	2000
9	0	0	0	0	0	0	0	909		1138	1202	1331	1440	1529	2000	2000	2000	2000	2000	2000	2000
10	0	0	0	0	0	0	0	821	862		1091	1232	1323	1428	2000	2000	2000	2000	2000	2000	2000
11	0	0	0	0	0	0	0	661	798	909		1119	1244	1333	2000	2000	2000	2000	2000	2000	2000
12	0	0	0	0	0	0	0	579	669	768	881		1120	1229	2000	2000	2000	2000	2000	2000	2000
13	0	0	0	0	0	0	0	454	560	677	756	880		1146	2000	2000	2000	2000	2000	2000	2000
14	0	0	0	0	0	0	0	386	471	572	667	771	854		2000	2000	2000	2000	2000	2000	2000
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0		1110	1268	1328	1420	1519	1594
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	890		1099	1251	1338	1396	1498
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	732	901		1130	1252	1345	1433
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	672	749	870		1082	1234	1342
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	580	662	748	918		1116	1180
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	481	604	655	766	884		1119
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	406	502	567	658	820	881	
	3800	4507	5312	5973	6734	7463	8211	17810	18451	19243	19936	20754	21527	22279	31761	32528	33207	34051	34796	35491	36166
	1.90	2.25	2.66	2.99	3.37	3.73	4.11	8.91	9.23	9.62	9.97	10.38	10.76	11.14	15.88	16.26	16.60	17.03	17.40	17.75	18.08

Key:

true expected value
item number

item number

choice matrix (number of times
the column item was preferred
to the row item)

aggregate preference score
mean preference score

We repeated these simulations with the middle cluster randomly spaced, instead of evenly spaced between the remaining two clusters, and with the items randomly spaced within the clusters. For example, one randomization of the middle cluster location produced clusters centered at Vs of 2, 16, and 20, such that the middle cluster is closer to the upper cluster than to the lower (fig. 3-5). At lower levels of Δ , the uneven spacing of the clusters interferes with the ability of MPSs to produce a true interval scale metric of the Vs. At $\Delta = 1$, although the within-cluster scales are almost perfectly linear (fig. 3-5), the overall relationship is not ($R^2 = 0.91$) (table 3-4, fig. 3-5). Because the middle cluster is closer to the upper cluster, the MPSs of the items of the middle cluster under-estimate the true values of the items. The MPSs under-estimate the Vs of the items in the middle cluster because at $\Delta = 1$, there is no interaction between the lower and middle clusters but some interaction between the middle and upper clusters (table 3-6).³⁵ The relation of MPS to V improves at $\Delta = 2$ ($R^2 = 0.94$) and is nearly linear by $\Delta = 6$ ($R^2 > 0.99$, table 3-4).

As with evenly spaced clusters, the three within-group MPS scales all have slopes of about 0.4 at $\Delta = 1$ and greater slopes at higher levels of Δ . The T and M scales offer modest improvements by reducing the overall non-linearity (table 3-4).

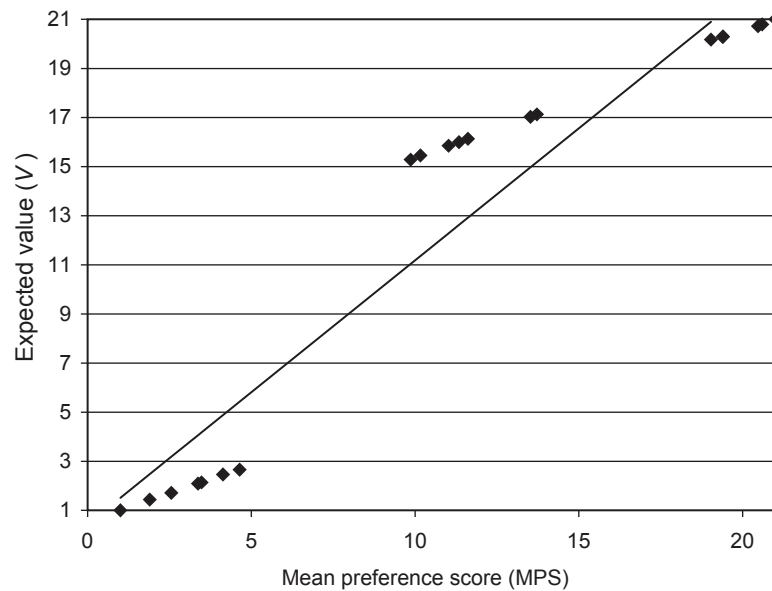


Figure 3-5. Relation of V to MPS, three randomly spaced clusters and randomly spaced V s within clusters, 21 items, $\Delta = 1$.

Table 3-6. Aggregate choice matrix, V s in three unevenly spaced clusters, unevenly spaced within clusters, $\Delta = 1$, 2,000 cases.

	1.00	1.44	1.72	2.09	2.14	2.46	2.66	15.28	15.46	15.85	16.00	16.13	17.02	17.13	20.17	20.29	20.30	20.73	20.79	20.99	21.00
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1		1272	1361	1613	1582	1702	1756	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
2	728		1146	1358	1417	1535	1636	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
3	639	854		1236	1194	1349	1471	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
4	387	642	764		1037	1224	1359	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
5	418	583	806	963		1176	1287	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
6	298	465	651	776	824		1154	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
7	244	364	529	641	713	846		2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000
8	0	0	0	0	0	0	0		1068	1270	1413	1458	1785	1818	1999	1999	2000	2000	1999	2000	2000
9	0	0	0	0	0	0	0	932		1258	1301	1344	1728	1765	1998	2000	2000	2000	2000	2000	2000
10	0	0	0	0	0	0	0	730	742		1069	1143	1642	1601	1994	2000	1998	1999	1999	2000	2000
11	0	0	0	0	0	0	0	587	699	931		1086	1539	1571	1994	1998	1996	1999	2000	1999	2000
12	0	0	0	0	0	0	0	542	656	857	914		1469	1529	1997	1996	1998	1998	2000	2000	1998
13	0	0	0	0	0	0	0	215	272	358	461	531		1086	1971	1989	1980	1992	1992	1996	1993
14	0	0	0	0	0	0	0	182	235	399	429	471	914		1967	1969	1975	1988	1994	1997	1997
15	0	0	0	0	0	0	0	1	2	6	6	3	29	33		1079	1079	1348	1352	1447	1449
16	0	0	0	0	0	0	0	1	0	0	2	4	11	31	921		989	1255	1253	1375	1408
17	0	0	0	0	0	0	0	0	0	2	4	2	20	25	921	1011		1196	1278	1389	1381
18	0	0	0	0	0	0	0	0	0	1	1	2	8	12	652	745	804		1011	1135	1118
19	0	0	0	0	0	0	0	1	0	1	0	0	8	6	648	747	722	989		1065	1095
20	0	0	0	0	0	0	0	0	0	0	1	0	4	3	553	625	611	865	935		1027
21	0	0	0	0	0	0	0	0	0	0	0	2	7	3	551	592	619	882	905	973	
	2714	4180	5257	6587	6767	7832	8663	17181	17674	19083	19601	20046	23164	23483	32166	32750	32771	34511	34718	35376	35466
	1.36	2.09	2.63	3.29	3.38	3.92	4.33	8.60	8.84	9.54	9.80	10.02	11.58	11.74	16.08	16.38	16.39	17.26	17.36	17.69	17.73

Key:

true expected value
item number

choice matrix (number of times
the column item was preferred
to the row item)

aggregate preference score
mean preference score

Heterogeneous Discriminal Dispersions

We have found in experiments with real respondents that preference score variance is smallest for choices involving dollar magnitudes, mid-range for choices between private goods, and largest for choices between public goods. We expect that those types of dispersion differences apply as well to the underlying V s of the items involved in the choices.³⁶ We have also observed a wide range across people in the consistency with which they judge items in paired comparison experiments, and we therefore expect that discriminial dispersions are heterogeneous across respondents. In this section we introduce heterogeneity in dispersion across items and across respondents. Thus, we remove “identically” distributed from the IID condition.

We present results for evenly spaced, randomly spaced, and clustered V s. As above, the V s of the 21 items range from 1 to 21. For these simulations, the respondents are separated into three sets of 2,000 respondents each. Within each set, each of the 21 items is randomly assigned a Δ of 1, 2, 3, 4, or 5. Combining the three sets of responses, we have choices for 6,000 virtual respondents, each set containing 2,000 identical respondents whose Δ s differ by item, with each of the three sets using a unique random assignment of the Δ s to the items.

Evenly Spaced V s

With evenly spaced V s, the MPSs perfectly preserve the ordinal relationship among the V s and comprise a nearly perfect interval scale, as was the case for IID discriminial dispersions. The R^2 between V s and MPSs exceeds 0.99 (table 3-7). The relationship is systematically non-linear, but the departure from linearity is trivial. In our simulation, estimation of psychometric scales required arbitrary adjustment of only three vacant cells in the aggregate choice matrix. The psychometric scales, T and M , also achieve R^2 s above 0.99 (table 3-7).

Table 3-7. R^2 for scale values versus V s, heterogeneous Δ from 1 to 5, 6000 cases.

Scale	V s	
	Evenly spaced	Randomly spaced
MPS	.9967	.9742
T	.9966	.9960
M	.9948	.9937

Randomly Spaced Vs

With randomly spaced V s (using the same random assignment as fig. 3-2), the relationship between V and MPS is moderately non-linear ($R^2 = 0.97$, table 3-7). The two psychometric scaling procedures achieve R^2 s with V above 0.99 (table 3-7), as in the case of IID discriminial dispersions. In our simulation, estimation of the psychometric scales required arbitrary adjustment of only five vacant cells.

Vs in Three Clusters

With V s in three clusters, when the three clusters are evenly spaced (centered at V s of 2, 11, and 20, as in fig. 3-4), the MPS scale and the psychometric scales very closely represent the V s (the overall R^2 s > 0.99) and all three clusters fall on the overall regression line. For MPSs, the within-cluster slope is almost identical to the overall slope in both end clusters (slopes of 0.980 and 1.002), but the within-cluster slope of the middle cluster is flatter than the overall slope (0.871).³⁷ This comparison suggests that when V s are evenly spaced within evenly spaced clusters, when the spacing within cluster is small relative to the spacing between clusters, and when discriminial dispersions are heterogeneous across items and respondents and averaging $\Delta = 3$, the MPSs are very similar to the psychometric scales and are likely to preserve significant interval relationships. In other words, the elevated mean Δ allows sufficient interaction between clusters to ameliorate the effect of the clustering on the estimation of V .

When the middle cluster is not evenly spaced between the two end clusters, but rather is centered at $V = 16$ (as in fig. 3-5), the clusters do not fall on the overall regression line. Moving the middle cluster away from the center of the V scale causes some non-linearity in the overall relation of V s to MPSs ($R^2 = 0.97$). The MPSs over-estimate the size of the interval between the middle and upper cluster, relative to the interval between the lower and upper cluster, as in the case of IID discriminial dispersions. The psychometric scaling procedures remove the distortion ($R^2 > 0.99$), which is notable given the violation of the assumption of homogeneity of variance, but they introduce some confusion (scatter of points) within clusters. In our simulation, estimation of the T and M scales required arbitrary adjustment of 13 vacant cells in the aggregate choice matrix.

Vs in Two Clusters

If the V s are grouped into two widely spaced clusters (fig. 3-6) and the V s of the items are evenly spaced within the clusters, the MPSs and the psychometric scales correlate strongly with the V s between clusters ($R^2 > 0.99$) and also within clusters. The slopes within the two clusters are about 0.62, compared with the overall slope of about 1.0 (fig. 3-6). When the V s are randomly spaced within clusters, results are similar to those of V s evenly spaced within the clusters. Thus, grouping into two clusters does not appear to present a problem (note, however, that we did not investigate the effects of altering the groupings so that some clusters have many more items than another).

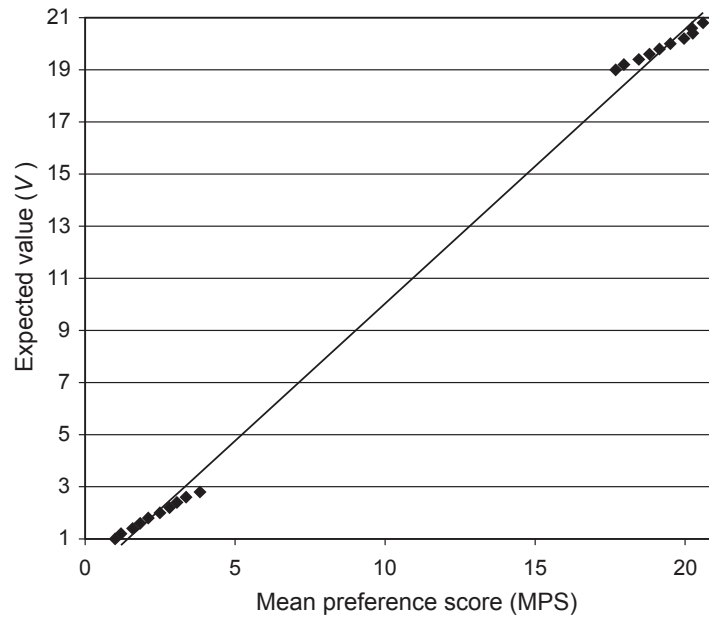


Figure 3-6. Relation of V to MPS, two clusters with evenly spaced values within cluster, 21 items, heterogeneous Δ .

Gross Heterogeneity of V s and Discriminal Dispersions

People do not necessarily agree with each other in their preferences. Even if they order items identically, their interval sizes are likely to differ. To test the effect of gross heterogeneity in both V and Δ on the various scaling options, we combined the results of 36 different simulations mentioned above. The aggregate choice matrix is the result of 72,000 virtual respondents making 15,120,000 choices among 21 items with a V range from 1 to 21 under a hodge-podge of different conditions. The only underlying fixed conditions are normally distributed discriminal dispersions and a constant ordinal relationship among the V s of the 21 items.

Under these conditions of gross heterogeneity, the MPSs, as well as the psychometric scales, comprise a nearly perfect interval scale of the average of the V s used in the several simulations (R^2 s > 0.99). Based on this analysis, we hypothesize that if the heterogeneity in the sample is random and ordinality of V s is maintained across respondents, the combined effect of intra- and inter-personal heterogeneity of V s and dispersions will be that the MPSs converge to a nearly perfect interval scale of the average of all the V s in the sample as the sample size approaches infinity. In a paired comparison experiment with a group of heterogeneous real people, however, we might expect violation of the assumptions of V range and ordinality as well as normality of dispersion. We have not tested the effects of relaxing those assumptions.

Effect of Sample Size

Throughout this chapter we have maintained a sample size of at least 2,000 virtual respondents in order to eliminate any effect of sample size. We now examine how closely scale values match V s at more typical sample sizes. For this comparison, we use a realistic formulation: 21 items with randomly spaced V s and heterogeneous discriminial dispersions (Δ varies from 1 to 5 across the items).

Figure 3-7 shows the correlation (R^2 , not adjusted R^2) of V to scale value (MPS, T , M) for sample sizes ranging from one to 100. Each data point represents an average of 30 R^2 s from 30 separate runs of the simulation program.³⁸ With a sample size of 2, the three scaling procedures are equivalent, producing an R^2 of about 0.94. The improvement in R^2 is rapid for increases in sample size up to about 10, reaching an R^2 of about 0.965. Up to this point there is little difference among the three scaling procedures, but beyond this point the improvement over MPSs gained with the more involved procedures is notable, although not large. The R^2 s for MPSs stabilize at about 0.97 by a sample size of 20, whereas the R^2 s for the other two procedures climb to about 0.98 by a sample size of 60. As reported earlier (table 3-7), with an unlimited sample size, the R^2 s reach 0.97 for MPS and 0.99 for the psychometric scales.

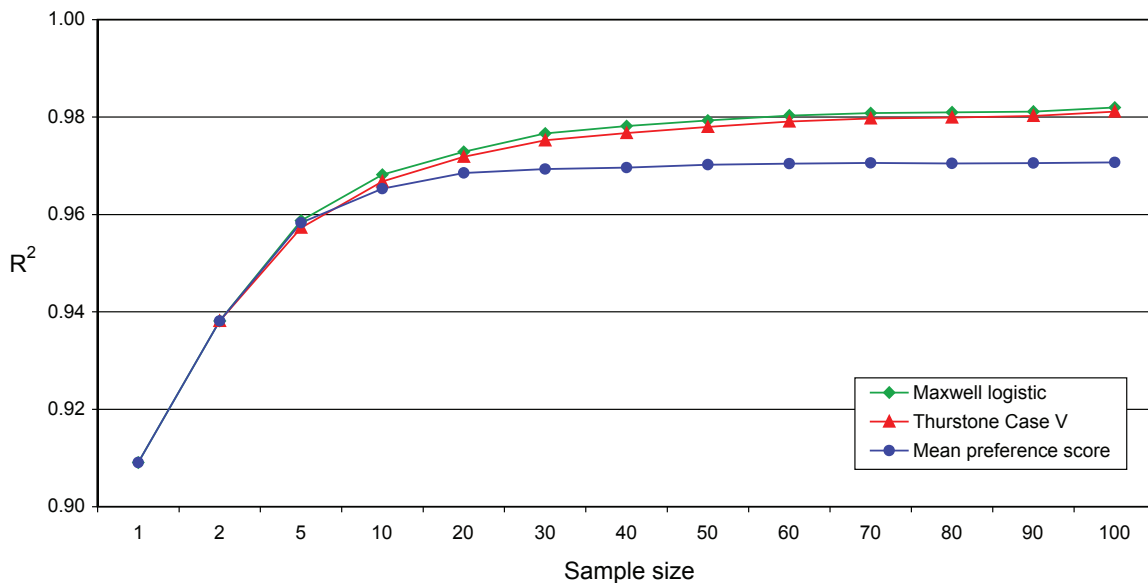


Figure 3-7. Correlations of scale value to V , 21 items, randomly spaced V s, heterogeneous Δ .

Summary and Conclusions

Our simulations, which assume that dispersions are normally distributed and that all respondents are homogeneous in their preferences (their V s), support at least the following four conclusions. First, under the best of conditions—IID discriminial dispersions, evenly spaced V s, and a very large sample of respondents—MPSs are a nearly perfect representation of V s no matter what the variance of the dispersions. The scale rotates as the dispersion, modeled in terms of Δ , increases such that the MPS intervals become less sensitive to the V intervals, but linearity persists. Psychometric scaling procedures such as Thurstone's Case V (T) and Maxwell's logistic (M) offer only trivial improvements over MPSs.

Second, under more realistic conditions—discriminal dispersions that are independently but not identically distributed across the items, with V s that are randomly spaced across the preference continuum—MPSs tend to be ordinal but not quite linearly related to the V s. The non-linearity is minor, however, and not sufficient to invalidate use of the MPSs as an estimate of an interval scale of the V s if the sample of respondents is large. In our simulations, the psychometric scaling procedures improved the relation of scale value of V to $R^2 > 0.99$ compared with $R^2 = 0.97$ for MPSs.

Third, if the V s are grouped in randomly spaced clusters and distributed randomly within clusters, and if the variances of the discriminial dispersions differ substantially across items (Δ ranging from 1 to 5 in our simulations of three clusters), the expected MPSs do not correctly represent the between-cluster or within-cluster intervals. Psychometric scales largely correct for the between-cluster non-linearity, but neither MPSs nor the psychometric scales perform particularly well within clusters.

Fourth, the ability of scale values to represent the V s increases dramatically up to a sample of about 10 respondents and increases less so as the sample size continues to increase. Overall, there is a diminishing marginal utility to increasing sample size. Up to a sample of about 10, all three scaling procedures perform equally. Above about 10, the Thurstone and Maxwell scales outperform MPSs. Increasing the sample size beyond 30 offers virtually no improvement in the MPS scale, and increasing the sample size beyond 60 offers little improvement in the T or M scales. In sum, differences between the T and M scales are trivial, but in some situations the T and M scales can provide a slight improvement over MPSs.

Based on the simulations, grouping of items into clusters of V s appears to be a serious concern. However, it must be remembered that the simulations are of a homogeneous sample of respondents. In any actual sample of individuals, even individuals of identical socio-economic characteristics, preferences are likely to be heterogeneous. Such heterogeneity will naturally produce a greater degree of overlap between clusters in the aggregate choice matrix than was found for the strict clustering depicted in figure 3-5. Items may be clustered for some respondents and spread out for others, and items may cluster differently for different respondents. In addition, as reported in Chapter 5, most respondents exhibit a Δ greater than the $\Delta = 1$ of figures 3-5 and 3-6, thereby further increasing interaction between clusters.

In summary, it can be concluded that (1) preference scores are, in general, about as good as more complex scaling methods and simpler to calculate and use; (2) one should strive for a sample size of at least 30 and strictly avoid a sample size below 10; and (3) one should avoid clustered V 's and strive for evenly spaced, or at least randomly spaced, V 's, although the natural variability among respondents and the magnitude of typical discriminial dispersions will tend to ameliorate the effect of any clustering. Avoiding clusters may require a multiple stage application of paired comparisons, with the initial stage used to get a first impression of V spacing, which is then used to design the choice set so as to approximate an even spacing of items.

Chapter 4. Effectiveness of the Double-Sort Procedure

Introduction

Let us assume, as in Chapter 1, that a respondent's paired comparison choices obey Thurstone's (1927a) law of comparative judgment, which posits that the respondent chooses between two items, say i and j , by taking random draws from discriminial dispersions about expected outcomes, or V s. If, over many independent choices between i and j , the respondent would prefer j more often than i , the respondent's expected preference order is $j \rightarrow i$.³⁹ It would be useful in an application of the paired comparison method, wherein a person responds only once to each pair, to identify those choices that are *consistent* with the person's preference order and those that are not. Unfortunately, determining which choices are inconsistent requires *a priori* knowledge of the very thing the experiment is designed to uncover: the expected preference order. Because the expected values of the preference functions (the V s) are not known, the consistency of a respondent's individual responses cannot be observed directly.⁴⁰

A successful detection of inconsistent choices would have two important uses. First, were it to be accomplished quickly—that is, immediately after the full set of pairs had been presented to a respondent of a paired comparison survey, as is possible if the survey is administered by computer—it would allow the inconsistent pairs to be presented again to the respondent during the same survey. Such a retesting would help determine whether the detected inconsistent responses were just a mistake or were indeed the choices the respondent intended to make. Second, detection of inconsistent choices would allow the experimenter to determine whether inconsistent choices are more likely in some circumstances than others. For example, degree of inconsistency may vary by the type of item being compared, the relative placement of the choice within the full set of choices presented (such as an early choice versus a late choice), or the amount of time respondents take to make their choices.

The DS Procedure

In this chapter, we examine the extent to which the double-sort (DS) procedure circumvents the problem of unknown V s and provides useful information about inconsistent choices. As described earlier, the DS procedure arranges the columns of the choice matrix in order of increasing preference score, left to right, and arranges the rows of the matrix in order of increasing row sum, bottom to top. Such a reordering of the matrix produces a double-sorted matrix with mostly “1”s above the principal diagonal and mostly “0”s below, and always preserves the original (raw matrix) order of items that are tied in preference score (fig. 4-1).

The DS procedure accepts the vector of preference scores as an estimate of the respondent's expected preference order among the items in the set and identifies,

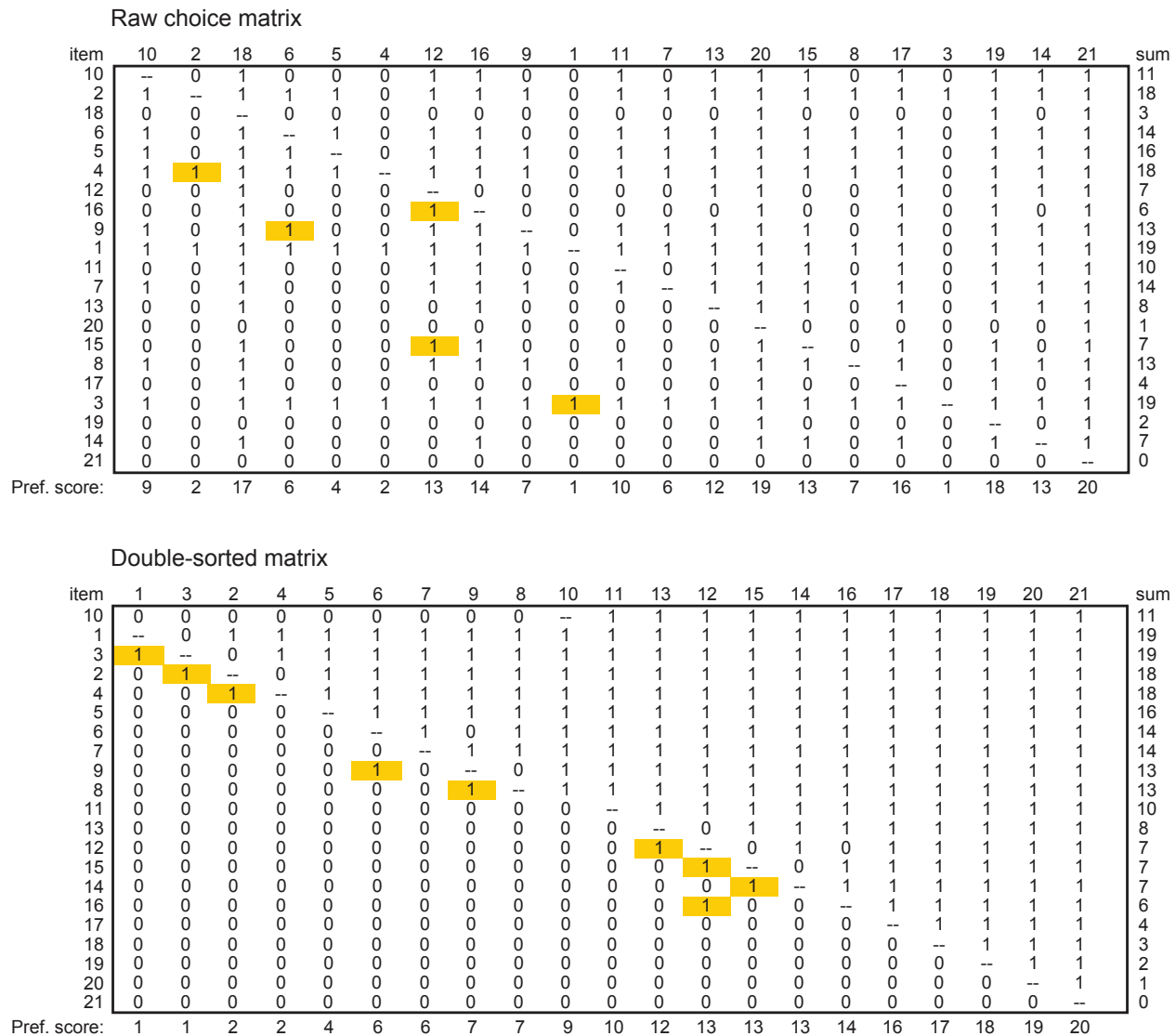


Figure 4-1. Hypothetical choice matrix with five inconsistent choices (highlighted in the raw choice matrix) and nine reported inconsistent choices in the double-sorted matrix. Item numbers, listed along the top and to the left of the matrices, indicate the Vs.

by a “1” below the principal diagonal, the choices that are likely to be inconsistent with those preference scores. A “reported” inconsistent choice is one that the procedure identifies as inconsistent with the preference order implied by the preference scores. We argue that, with the exception of items involved in preference score ties, the choices identified by the DS procedure as inconsistent are the best estimate of the set most likely to be inconsistent with the respondent’s expected preference order. However, as seen in detail below, the procedure is imprecise. A “reported” inconsistent choice is sometimes actually a consistent choice and a reported consistent choice is sometimes actually an inconsistent choice.

To demonstrate the workings of the DS procedure and determine whether the procedure may serve the uses listed above, we used the PCS model to determine the expected results of using the DS procedure, assuming normal discriminial dispersions. We also gauged the usefulness of the DS procedure by examining evidence from actual (in other words, real-people) experiments.

Using the PCS model, we imposed a preference order by specifying the V s, introduced inconsistent choices, and then observed whether the DS procedure found the inconsistent choices. We used three criteria to evaluate the effectiveness of the DS procedure: (1) choices correctly reported as inconsistent are “found,” (2) choices reported as consistent that actually are inconsistent are “not-found,” and (3) choices reported as inconsistent that actually are consistent are “bad-found.” If $N(\)$ indicates “number of” and A indicates an actual inconsistent choice, then:

$$\begin{aligned} N(\text{reported}) &= N(\text{found}) + N(\text{bad-found}), \\ N(A) &= N(\text{found}) + N(\text{not-found}) = N(\text{reported}) + N(\text{not-found}) - N(\text{bad-found}), \\ N(A) - N(\text{reported}) &= N(\text{not-found}) - N(\text{bad-found}). \end{aligned}$$

To gain a sense of how the PCS model is used to examine the effectiveness of the DS procedure, consider figure 4-1, which shows a matrix of 21 items with the V s set equal to the item numbers. The top matrix of figure 4.1 is the raw choice matrix, with the items randomly ordered. The bottom matrix is the double-sorted matrix. Given that the item numbers indicate the V s of the items, there are five inconsistent choices in the raw choice matrix, indicated by the shaded cells. In this particular case, the inconsistent choices produce eight circular triads and a coefficient of internal reliability (ζ) of 0.98. As seen more easily in the double-sorted matrix, the circular triads produce four two-way ties in preference score and one three-way tie. The order of tied items in the double-sorted matrix remains as it was in the raw choice matrix. Nine choices are reported as inconsistent in the double-sorted matrix, indicated by the “1”s below the principal diagonal. In this example, five of the “1”s represent the original five inconsistent choices (they are found), the other four are incorrectly identified (they are bad-found), and there are no not-found inconsistent choices.

Simulation Results

We use the simulation model for two purposes. First, we explore some issues regarding how to set up a paired comparison experiment so that the DS procedure will be as effective as possible. For this purpose, we examine three orderings of the V s in the raw choice matrix (ascending, random, and descending) and explore two spacings of the V s (evenly spaced and randomly spaced). Second, we examine in more detail how the DS procedure identifies inconsistent choices, and how the performance of the procedure varies with changes in the discriminial dispersions about the V s of the items. We ask if it makes a difference whether the discriminial dispersions are identical. To do this, we vary the magnitude of the standard deviation of the discriminial dispersions (Δ) from 1 to 10 for IID normal dispersions and examine one option with heterogeneous dispersions, wherein a Δ from 1 to 5 was randomly assigned to each item, resulting in an average Δ of about 3. This analysis helps determine whether the DS procedure is biased, or in other words, whether it can be employed for the uses listed at the beginning of this chapter—to identify inconsistent choices for retesting and to test for differences in consistency by type of item, sequence of presentation, and time

taken to respond. Addressing this second purpose essentially asks whether the procedure adequately detects the differences in Δ among the items (in other words, whether a difference in Δ across the items will result in differences in the incidence of inconsistent choices that are detected by the DS procedure).

All simulations use 21 items with V 's that range from 1 to 21. Each simulation was run with a homogeneous sample of 2,000 independent virtual respondents (in other words, 2,000 independent but identical individuals). Each virtual respondent is presented with all possible pairs of the 21 items, resulting in 210 choices per respondent for a total of 420,000 choices in each simulation.

To set the stage, we present the cumulative distributions of all choices as a function of preference score difference (PSD). We do this for the case of randomly spaced V 's with the items randomly ordered in the raw choice matrix (fig. 4-2), but other selections produce a similar result. As figure 4-2 shows, regardless of Δ , the choices are concentrated in the lower range of PSD, which reflects the assumption of normal discriminial dispersions. However, the degree of concentration varies by Δ , with the proportion of choices above a given PSD decreasing as Δ increases. For example, given $\Delta = 2$, 64 percent of the choices are between pairs with a $\text{PSD} \geq 5$, whereas with $\Delta = 10$, only 37 percent of the choices are between pairs with a $\text{PSD} \geq 5$.⁴¹ In other words, a smaller Δ (thus, a more narrow discriminial dispersion) leads to a lower number of inconsistent choices and thus to fewer preference score ties, producing a larger proportion of choices with high PSDs.

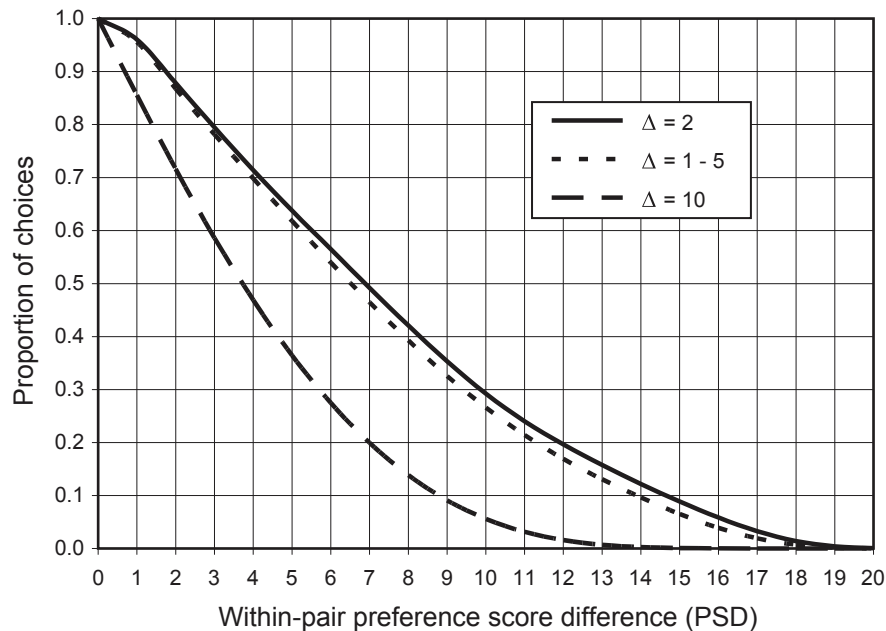


Figure 4-2. Reverse cumulative proportion of paired choices for three specifications of Δ (IID normal discriminial dispersions with $\Delta = 2$ or 10, and heterogeneous normal discriminial dispersions with Δ from 1 to 5), given randomly spaced V 's and a random raw matrix input order for the items.

Issues Regarding the Setup of the Paired Comparison Experiment

Effect of the order of the items in the raw choice matrix—Figure 4-3 shows, for evenly spaced V s in ascending, random, and descending input order and IID normal discriminial dispersions, the relationship between PSD and the proportion of the inconsistent choices correctly found by the DS procedure. The figure shows that V input order is irrelevant for $\text{PSD} \geq 1$, and that for choices with $\text{PSD} = 0$, the raw matrix input order makes a profound difference. When the input order is ascending, the DS procedure always correctly finds all the inconsistent choices between items with the same preference score, but when the input order is descending, the DS procedure never finds them. When the input order is random, the expected outcome of the DS procedure is to correctly find half of the inconsistent choices between items with tied preference scores.

Ignoring items with a $\text{PSD} = 0$, the DS procedure correctly found 60 percent of the inconsistent choices, with the portion found rising from 52 percent at $\text{PSD} = 1$ to nearly 100 percent by $\text{PSD} = 6$. When tied items are included, the DS procedure correctly found 69 percent, 58 percent, and 46 percent of the inconsistent choices for the ascending, random, and descending input orders, respectively.⁴² Because the number of choices varies inversely with PSD, the proportion correctly found depends most heavily on how well the DS procedure deals with choices between items with a small PSD.

Because a random input order is the most realistic case, subsequent simulations presented here employ a random input order.

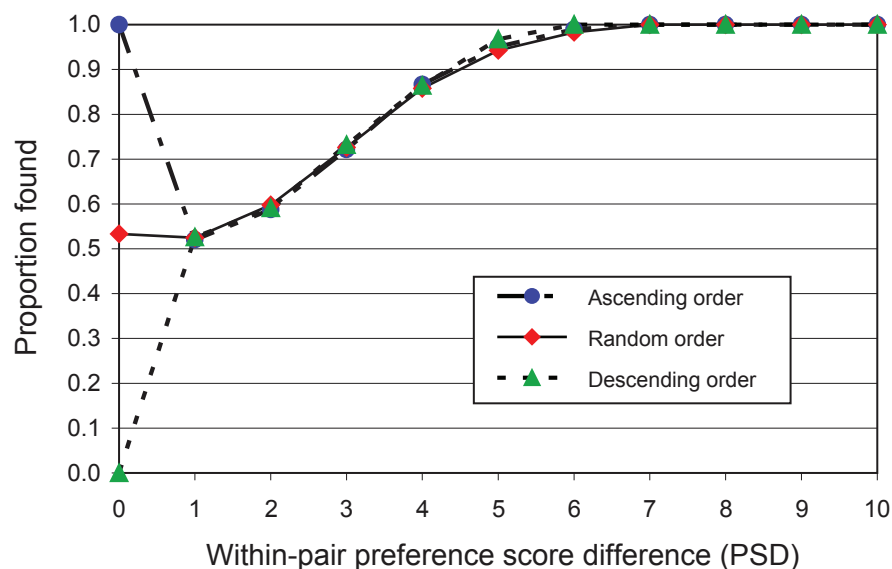


Figure 4-3. Proportion of inconsistent choices found by the DS procedure for three raw matrix input orders for the items, given evenly spaced V s and IID normal discriminial dispersions with $\Delta = 2$.

Effect of spacing of the V s—Figure 4-4 compares evenly and randomly spaced V s, again for items with IID normal discriminial dispersions of $\Delta = 2$. The randomly spaced V s are not integer values. The figure shows that the performance of the DS procedure declines slightly when the V s are randomly spaced as opposed to evenly spaced. This finding holds for ascending and descending input orders as well, and also for $\Delta > 2$. Given the random item input order and $\Delta = 2$ of figure 4-4, the DS procedure correctly found 53 percent of the inconsistent choices when the V s were randomly spaced, compared with 58 percent when they were evenly spaced (ties included).⁴³

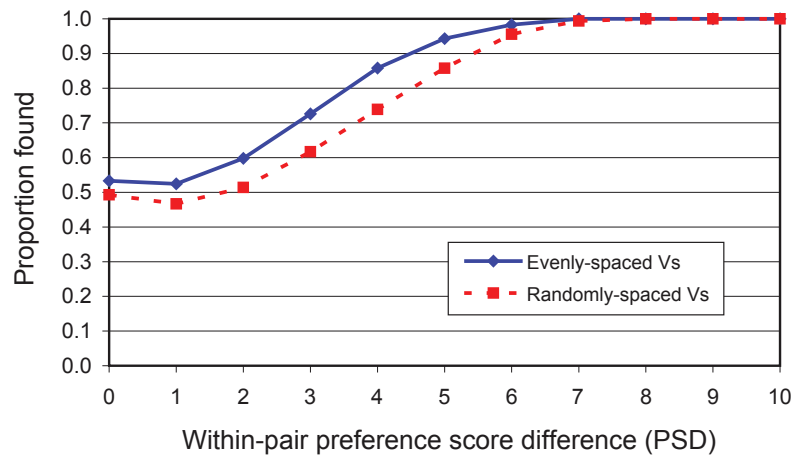


Figure 4-4. Proportion of inconsistent choices found by the DS procedure for evenly spaced and randomly spaced V s, given V s, IID normal discriminial dispersions with $\Delta = 2$, and a random row matrix input order for the items.

How Well Does the DS Procedure Work?

Reported versus true inconsistency—As described above, the DS procedure does not find all inconsistent choices, and it reports as inconsistent some choices that actually are consistent. Figure 4-5 depicts the relation between true and reported inconsistent choices as a function of PSD for the case of IID normal discriminial dispersions with $\Delta = 2$. Overall, 9 percent of the choices are inconsistent while only 7 percent are reported as inconsistent. The reported proportion underestimates the true proportion up to a PSD of 6, with the maximum discrepancy between true and reported inconsistencies at PSD = 1. The DS procedure presumably underestimates the true number of inconsistent choices because not all truly inconsistent choices actually cause circular triads. The curve for reported inconsistency appears to have a negative exponential functional form, whereas the true proportion appears to be a biased ogive function of PSD.

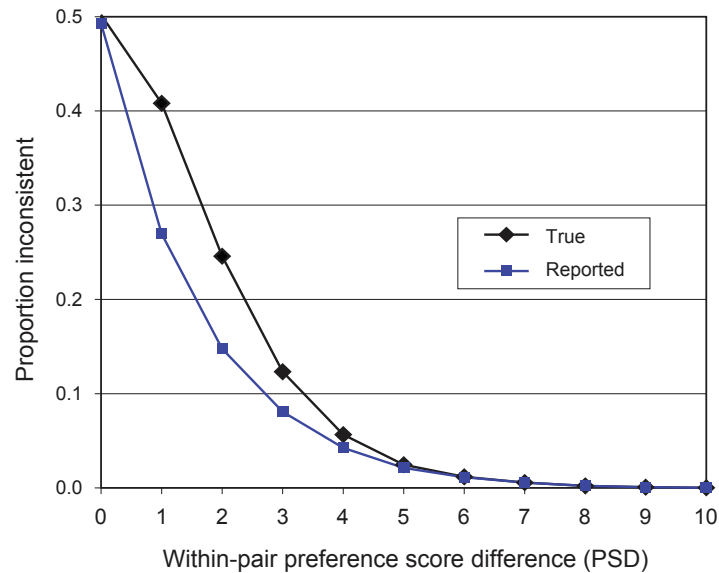


Figure 4-5. Proportions of all choices that are inconsistent and that are reported by the DS procedure as being inconsistent, given IID normal discriminial dispersions with $\Delta = 2$, randomly spaced Vs, and a random raw matrix input order for the items.

Effect of discriminial dispersion standard deviation (Δ)—Figure 4-6 shows the proportion of all 420,000 choices that are inconsistent versus the proportion that are reported as inconsistent given IID normal discriminial dispersions with alternative standard deviations ($\Delta = 1, 2, \dots, 10$) or heterogeneous normal dispersions (Δ ranging from 1 to 5 across the items). The figure shows three important findings. First, the number of inconsistent choices increases with Δ , as would be expected given the normal distributions of the discriminial dispersions. Second, the number of choices reported as inconsistent always underestimates the actual number of inconsistent choices (in other words, the number of choices not-found exceeds the number bad-found). The degree of under-reporting drops in percentage terms as Δ increases up to $\Delta = 8$. For example, the number of choices reported as inconsistent is 73 percent of the actual number inconsistent at $\Delta = 2$ and approximately 87 percent for Δ from 6 to 10. The result for heterogeneous discriminial dispersions with Δ from 1 to 5 is very similar to that of $\Delta = 3$.

Third, figure 4-6 shows that the relation of the proportion of actual inconsistent choices (IC) to the proportion of reported inconsistent choices (IC_r) is nearly linear across the range of Δ from 1 to 10. Linear regression reveals the following:

$$IC = 0.018 + 1.075 IC_r \quad (23)$$

Thus, given the conditions of figure 4-6, the true number of inconsistent choices can be estimated from the reported number regardless of the standard deviations of the discriminial dispersions of the respondents, up to Δ of at least 10.

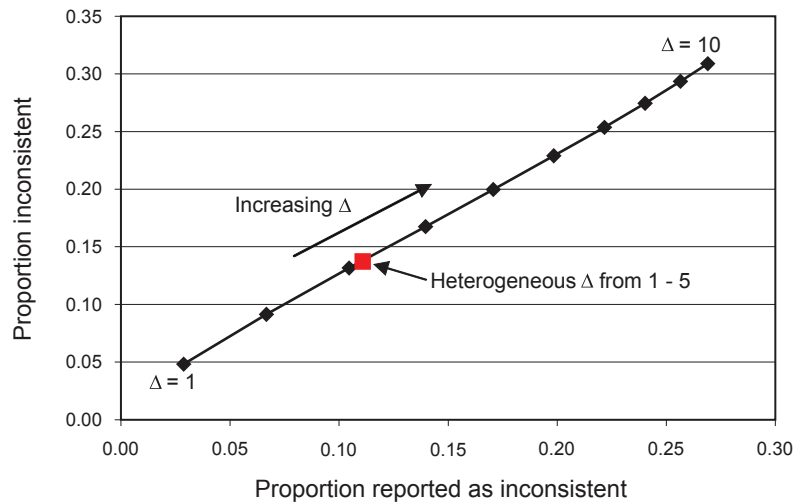


Figure 4-6. Proportions of all choices that are inconsistent and reported by the DS procedure as being inconsistent, for alternative specifications of Δ (IID normal discriminial dispersions with $\Delta = 1, 2, \dots, 10$, and heterogeneous normal discriminial dispersions with Δ from 1 to 5), given randomly spaced V s and a random raw matrix input order for the items.

Figure 4-7 shows the effect of alternative discriminial dispersions for the following three specifications of Δ : $\Delta = 2$ for all items, $\Delta = 10$ for all items, and a random assignment across items of Δ from 1 to 5. The found relation of proportion to PSD becomes straighter as Δ increases (reflecting the greater overlap of discriminial dispersions as Δ increases). The performance of the DS procedure in the heterogeneous case ($\Delta = 1 - 5$, which produces an average Δ of about 3) is, understandably, between that of the $\Delta = 2$ and $\Delta = 10$ cases. The DS procedure correctly found 53 percent of the inconsistent choices with $\Delta = 2$, 60 percent with Δ from 1 to 5, and 66 percent with $\Delta = 10$.

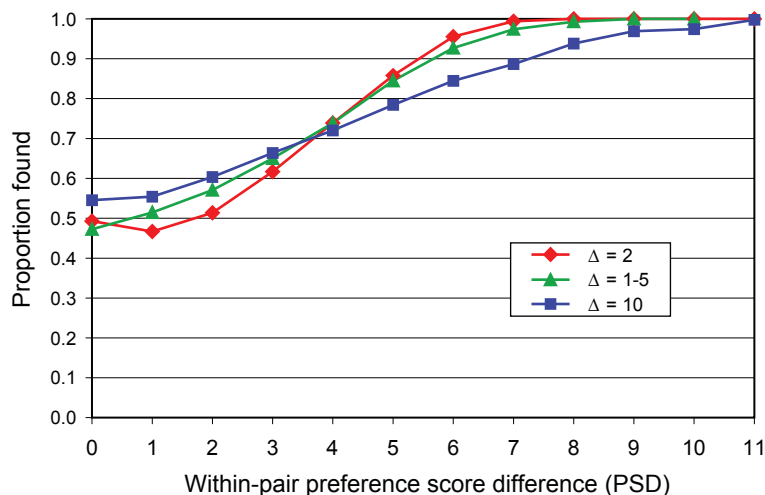


Figure 4-7. Proportion of inconsistent choices found by the DS procedure for three specifications of Δ , given randomly spaced integer V s, independent normal discriminial dispersions, and a random raw matrix input order for the items.

Reported versus found—Figure 4-8 shows the proportion of inconsistent choices that are reported as inconsistent along with the proportion that are correctly found. Also shown is the proportion of inconsistent choices that is bad-found, which is equal to the proportion reported minus the proportion found. At $PSD = 0$, the number of bad-found choices is about the same as the number found, both of which approximate 49 percent of the inconsistent choices. However, the number of bad-found choices drops quickly to 19 percent, 9 percent, and 4 percent at $PSD = 1, 2,$ and 3 , respectively. In other words, the selection of inconsistent choices by the DS procedure at $PSD = 0$ is essentially arbitrary (given the random input order of the items in the raw choice matrix), but the likelihood that a reported inconsistent choice is an actual inconsistent choice increases rapidly as PSD increases.

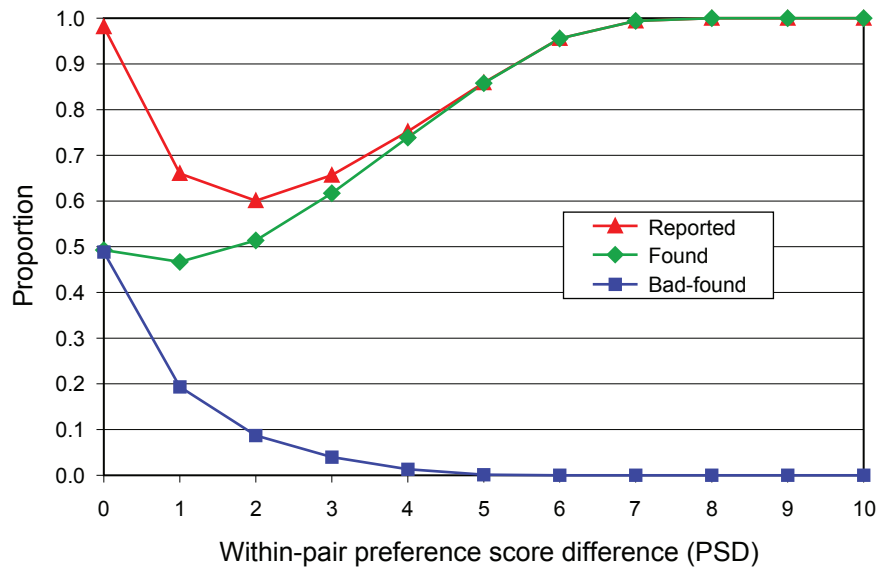


Figure 4-8. Proportion of inconsistent choices reported as inconsistent, found and bad-found by the DS procedure, given randomly spaced Vs, IID normal discriminial dispersions with $\Delta = 2$, and a random raw matrix input order for the items.

Figure 4-9 is similar to figure 4-8 except that it shows proportions of reported, not total, inconsistent choices that are found and bad-found. In a real-people experiment, the actual inconsistent choices, of course, are not known and the reported inconsistencies are all we have to go on. Found choices reach 94 percent of reported inconsistent choices by a PSD of only 3, and 98 percent by $PSD = 4$. Bad-found choices are correspondingly low.

Figure 4-10 presents proportions found and bad-found versus Δ , showing that although the proportion found is not very sensitive to Δ , the proportion increases as Δ increases up to $\Delta = 5$ and then drops slightly as Δ rises further.

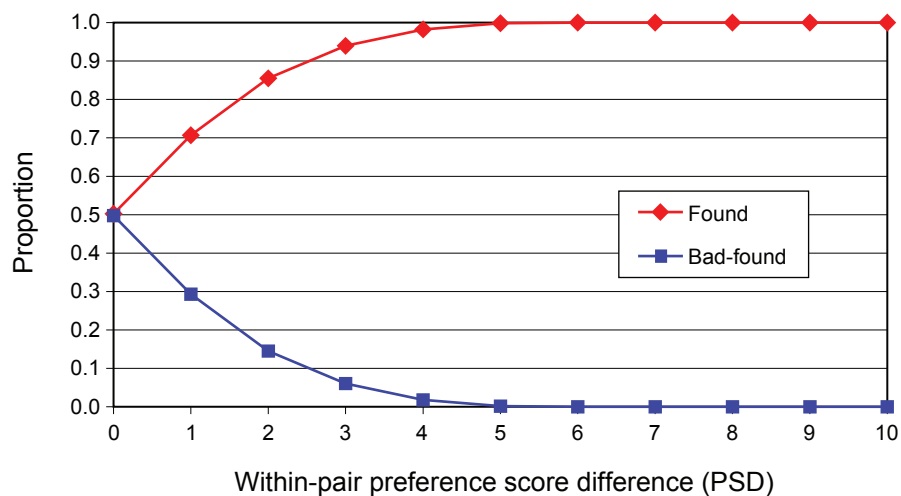


Figure 4-9. Proportion of reported inconsistent choices found and bad-found by the DS procedure, given randomly spaced V s, IID normal discriminial dispersions with $\Delta = 2$, and a random raw matrix input order for the items.

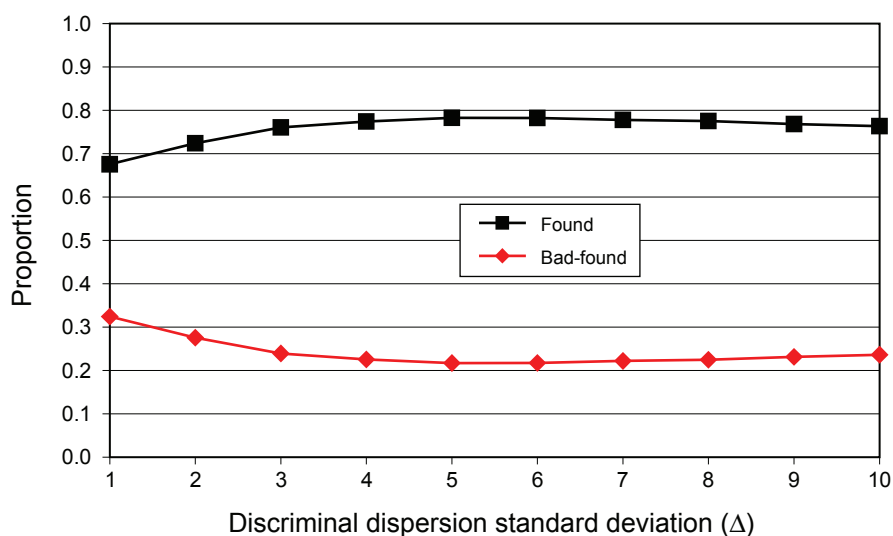


Figure 4-10. Proportion of reported inconsistent choices found and bad-found by the DS procedure at alternative levels of Δ , given randomly spaced V s, IID normal discriminial dispersions, and a random raw matrix input order for the items.

Summary

We have learned through numerical exercises, simulation modeling, and theoretical reasoning that the likelihood that a reported inconsistency is a true inconsistency increases as (1) the real within-pair difference between the V s increases, (2) PSD increases, (3) the spacing of V s becomes more uniform, and (4) the order of the items in the raw choice matrix more closely reflects increasing V , other things being equal.

We recommend that if the DS procedure is used in a paired comparison experiment to identify inconsistent choices, clusters of similarly valued items should be avoided and the order of the items in the individual respondent's raw choice matrix should be random unless there is *a priori* knowledge of the expected preference order. If the expected preference order is known, even approximately, that information should be used. If the input order is random and $\Delta = 2$, the overall expected proportion of inconsistent choices found should be at least 0.50 and the proportion of reported inconsistent choices that are correctly found should be about 0.72.

Figure 4-11 reports on the success of the DS procedure as a function of PSD for the recommended situation (randomly spaced V s, with the items in random order in the raw choice matrix) and for IID normal discriminial dispersions with a realistic standard deviation ($\Delta = 2$). The four curves of the figure are cumulative proportions, two with negative slopes reflecting accumulation from right to left, and the other two with positive slopes. The curves with negative slopes depict proportions of all choices and of the inconsistent choices, showing, for example, that 88 percent of all the choices, but only 42 percent of the inconsistent choices, are between items with a $\text{PSD} \geq 2$. The curves with positive slopes depict proportions of choices found, showing, for example, that among choices for which $\text{PSD} \geq 2$, 60 percent of all inconsistent choices are found (thus 40 percent are not-found), and 91 percent of the reported inconsistent choices are found (thus 9 percent are bad-found).

We conclude the following from figure 4-11. First, the number of inconsistent choices is inversely proportional to PSD, such that 58 percent of all inconsistent choices are between item pairs of $\text{PSD} < 2$. Second, the success of the DS procedure increases rapidly as PSD increases. For example, for inconsistent choices for which $\text{PSD} \geq 1, 2, 3, 4$, and 5, the percentages found are 54, 60, 70, 81, and 91, respectively.

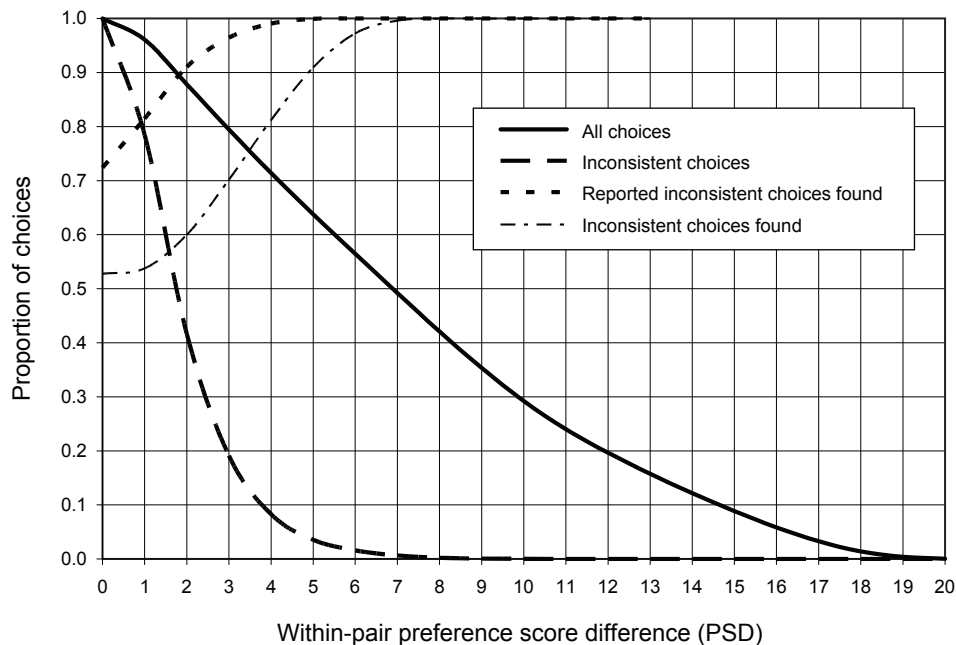


Figure 4-11. Cumulative proportions of choices by PSD, given randomly spaced V s, a random raw matrix input order for the items, and IID discriminial dispersions with $\Delta = 2$.

Third, the portion of reported inconsistent choices that are found (that are actually inconsistent choices) also increases with PSD, such that, for example, at $\text{PSD} \geq 1, 2, 3,$ and 4 the percentages found are 81, 91, 96, and 99, respectively (correspondingly, the percentages bad-found are 19, 9, 4, and 1). Clearly, the DS procedure is not perfect, but if the chance of a not-found and a bad-found does not interact with the measure of interest (such as sequence of the choice or type of item), we can rely on reported inconsistent choices as a measure of relative inconsistency.

Evidence From Real-People Experiments

It would seem at first that nothing could be said about the discriminial dispersions represented in the responses of real people in a paired comparison experiment. The response dispersions observed across people are likely to be a composite of within-person variation across items and between-person variation within items. Further, our real-people experiments, wherein the respondents made each choice only once (except for the retests), provide no way to distinguish between these two sources of variation. Comparison of our real-people results with our numerous simulation results, however, shows that what we will call the *effective* Δ of the dispersion in our real-people experiments is approximately 2. By *effective* Δ , we mean that the aggregate results of our real-people experiments are about the same as those produced by the simulations where all respondents are identical, each obeys the law of comparative judgment, and the Δ of each item is 2 relative to a scale on which the V s range between 1 and 21.

To support this claim, we present the results of several comparisons of real and simulated paired comparison experiments. First, figure 4-12 shows the proportion of all choices reported as inconsistent by the DS procedure as a function of PSD for both the real experiments and the simulations. The real results are from six independent experiments in which a total of 1,230 people made a total of 170,546 choices for choice sets ranging from 18 to 21 items (table 2-1). In every case, the sets consisted of a mixture of public goods, private goods, and sums of money, where for input to the PC program the goods were arranged in unknown order of V and the dollar magnitudes were always in ascending order of face value. The simulation results for each of 10 discriminial dispersion Δ s ranging from 1 to 10 are composites of six simulations across which 12,000 virtual respondents made a total of 2,520,000 choices. The simulations presented virtual respondents with 210 choices among 21 items where the V s ranged between 1 and 21. Each composite simulation result aggregates across evenly or randomly spaced V s that are in ascending, descending, or random input. Such a composite, we believe, best approximates the heterogeneity in a real experiment. The simulations did, however, comply strictly with the assumption of IID normal discriminial dispersion.

As figure 4-12 shows, the real person curve corresponds best to the simulation with a Δ of 2. The real curve falls slightly under the simulated $\Delta = 2$ curve for small PSDs and then climbs slightly above it as PSD increases beyond about 3. We speculate that the difference between the real and simulated curves arises from departure of individual behavior from the assumed IID normal dispersions, and from heterogeneity among the real respondents.

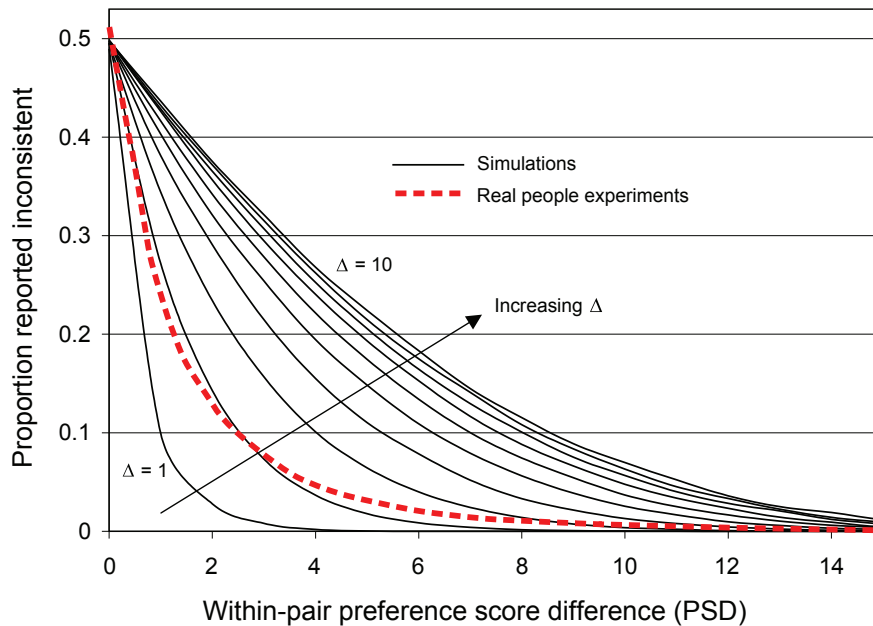


Figure 4-12. Proportion of choices reported as inconsistent versus PSD for real-people experiments and 10 simulations differing in Δ (1, 2, ..., 10). Real-people results are an aggregation of six experiments employing 1,230 respondents. Each simulation represents 12,000 virtual respondents with IID normal discriminial dispersions choosing among 21 items ranging in V from 1 to 21. Subgroups for each simulation respond to evenly or randomly spaced V s, with the items in ascending, random, or descending raw matrix input order.

Figure 4-13 compares the aggregate proportion of choices reported as consistent and inconsistent by the DS procedure from the six real-people experiments with true and reported proportion consistent and inconsistent from simulations for randomly spaced V s and IID normal discriminial dispersions with $\Delta = 2$. The proportions that are consistent range from about 0.5 at PSD = 0 to nearly 1.0 at a PSD = 7, whereas the proportions that are inconsistent fall in a mirror image from 0.5 to 0. The reported proportions from the real-people experiments are almost coincident with the reported proportions from the $\Delta = 2$ simulations. Over all choices, 0.931 of the real people choices were reported as consistent, compared with 0.934 for the $\Delta = 2$ simulations. These results suggest that from the simulations, the relationship that was found between the true proportion of choices that are inconsistent and the reported proportion can be used to estimate the true proportion of choices that are inconsistent in real-people experiments.

Figure 4-14 shows the aggregate proportion of choices reported as inconsistent as a function of the average coefficient of internal reliability (ζ) for a range of Δ . The simulations are the same as described above for figure 4-12. Figure 4-14 suggests two important conclusions. First, there is a strong relation between the expected proportion of choices that the DS procedure reports as inconsistent and the expected value of ζ . In other words, there is a strong relation between consistency and internal reliability. Second, the composite outcome from our six real-people experiments falls on the curve and is close to the simulation result for a Δ of 2.

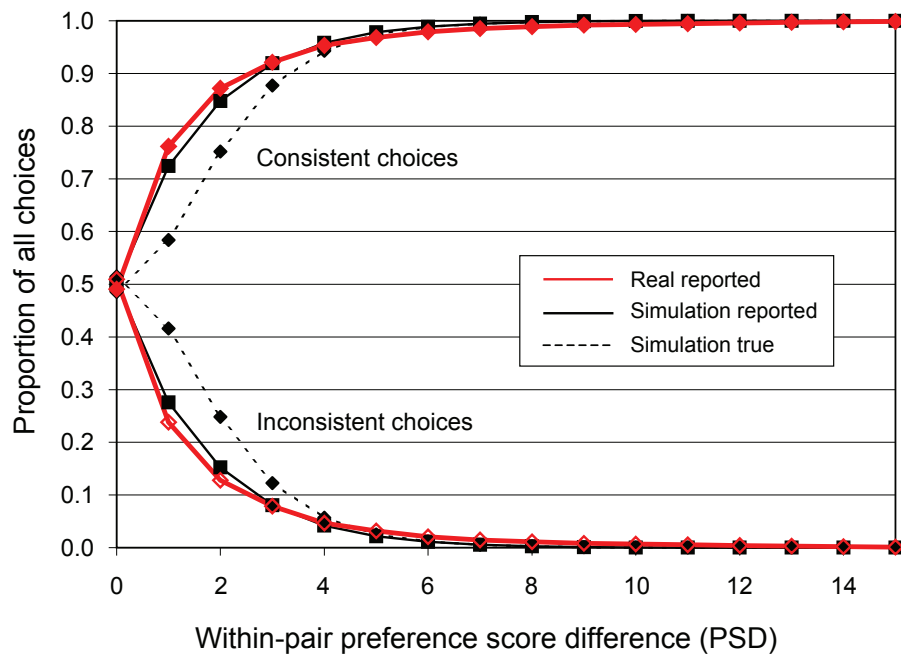


Figure 4-13. Proportion of choices reported as consistent and inconsistent versus PSD for real-people experiments and simulated results. Real-people results are an aggregation of six experiments employing 1,230 respondents. Simulations use 6,000 virtual respondents, IID normal discriminial dispersions with $\Delta = 2$, randomly spaced Vs ranging from 1 to 21, and a random raw matrix input order for the items. Also shown is the simulated proportion of choices that are actually consistent or inconsistent.

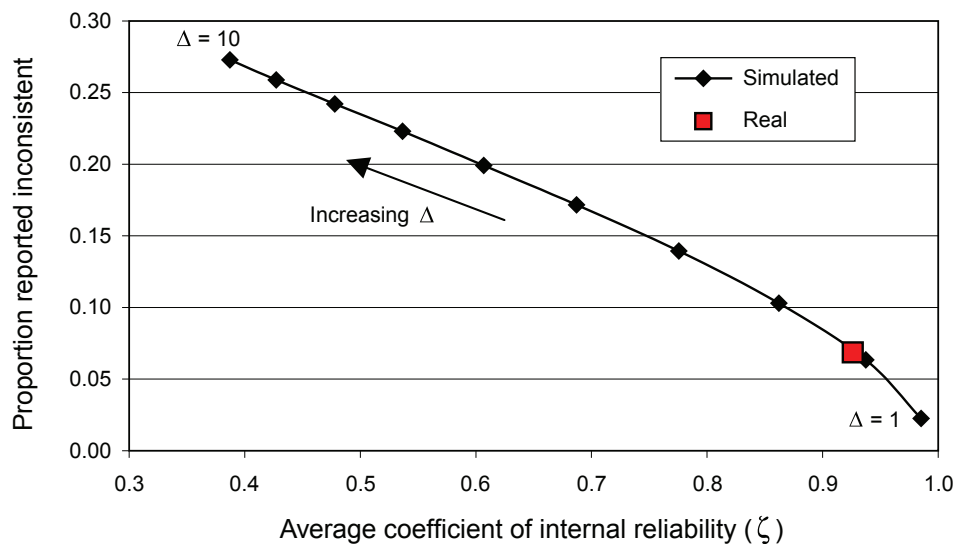


Figure 4-14. Proportion of choices reported as inconsistent versus coefficient of internal reliability for aggregate result of real-people experiments and 10 simulations at alternative Δ . Simulations use IID normal discriminial dispersions, Vs ranging from 1 to 21, and 12,000 virtual respondents at each of the 10 (1, 2, ..., 10) levels of Δ . Each simulation point represents a composite of six simulations (2000 respondents per simulation) using evenly or randomly spaced Vs and ascending, random, or descending raw input orders for the items.

Figure 4-15 plots ζ from the six real-people experiments as a function of the proportion of respondents in order of decreasing ζ . The results are shown for two conditions. In order to understand those two conditions, recall that the computer program that administered the real experiments used the DS procedure to identify reported inconsistent choices for each respondent in real time and required each respondent to repeat those choices at the end of the experiment without a perceptible break in the procedure. On the assumption that respondents discover or construct their values for difficult choices and/or correct mistakes as they progress through the randomly ordered set of choices, the experiment allows us to “correct” original choices by replacing the original choices with the choices made when the pairs were repeated.

Figure 4-15 shows that reversing the choices that were switched when retested improves the internal reliability of the respondents. The mean ζ increases from 0.93 to 0.98, and median ζ increases from 0.95 to 0.99. One interpretation of these results is again that respondents become more confident about their values as the experiment progresses through the sequence of pairs, such that when retested, respondents tend to correct mistakes or reverse choices about which they at first were uncertain or confused. This finding suggests that the DS procedure serves a useful purpose if the objective of the paired comparison exercise is to efficiently obtain a well-considered ordering of the items.

Figure 4-16 compares the reported number of inconsistent choices with the actual number of inconsistent choices in a similar manner to figure 4-6, but also includes a finding from the real-people experiments. Figure 4-16 is based on the composite results of numerous simulations under conditions of ascending, descending, and random V input order as well as even or random spacing of V s.

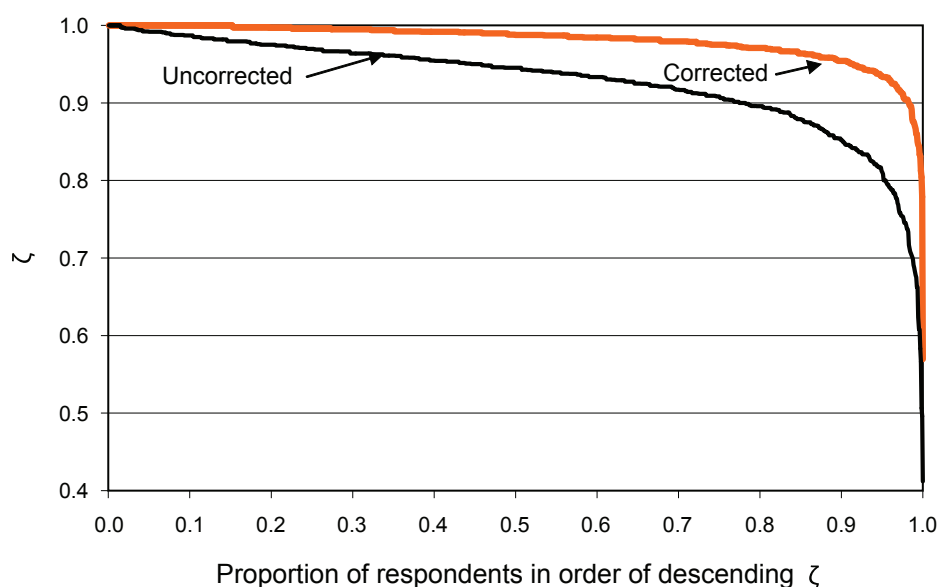


Figure 4-15. Retest-corrected versus uncorrected coefficient of internal reliability (ζ) for a composite of six real-people experiments.

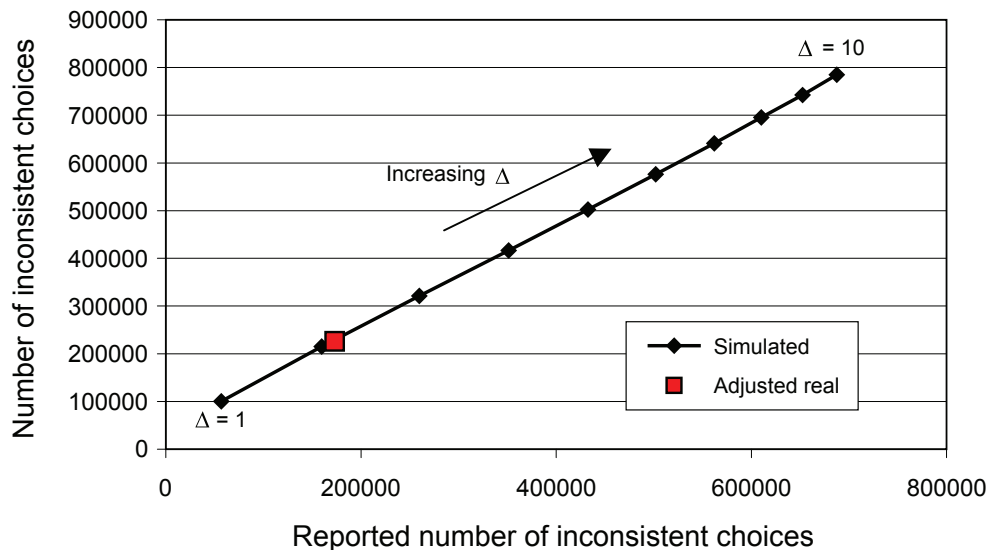


Figure 4-16. The number of inconsistent choices in real-people experiments versus simulated experiments. Simulations use IID normal discriminial dispersions, evenly and randomly spaced Vs ranging from 1 to 21, ascending, random, and descending raw matrix input orders for the items, and 12,000 virtual respondents at each of the 10 (1, 2, ..., 10) levels of Δ . The real-people result is scaled to correspond to the simulation sample size.

The total number of virtual choices at each level of Δ is 2,520,000. Regressing the true number of inconsistent choices (IC) on the reported number (IC_r) yields the following equation ($R^2 = 0.9998$):

$$IC = 40018 + 1.075 (IC_r). \quad (24)$$

Thus, the relationship between the actual number and the reported number of inconsistent choices is virtually linear. Accepting that the aforementioned theoretical assumptions (principally, IID normal discriminial dispersions) hold for real people, we can place a point in figure 4-16 representing the real-people studies. Our real-people studies produced 11,706 reported inconsistent choices from 170,546 total choices; thus, 6.86 percent of the choices were reported as inconsistent. Applying this percentage to the 2,520,000 simulated choices yields an estimate of 172,969 reported inconsistent choices. Substituting this estimate for reported inconsistent choices into equation 24 yields an estimate of 225,944 true inconsistent choices. As expected from figure 4-14, this point (172,969 reported versus 225,944 true) falls near $\Delta = 2$ on the simulated line (fig. 4-16).

From figure 4-16 we draw three conclusions. First, the real study results are comparable to simulation results based on IID normal discriminial dispersions where $\Delta = 2$ relative to an expected value range from 1 to 21. Second, the number of inconsistent choices reported in the real studies is approximately 77 percent of the true number ($172,969/225,944 = 0.766$). Third, we can predict the true number of inconsistent choices from the number reported by the DS procedure.

Conclusions

The DS procedure relies on preference scores as estimates of the real expected values of the items. This reliance causes error in the designation of individual inconsistent choices, resulting in inaccuracies in the estimate of the aggregate number of inconsistent choices, as well as in the specification of individual inconsistent choices. First, regarding aggregate estimates, the number of choices reported by the DS procedure as inconsistent always underestimates the actual number. At a realistic average Δ of 2, based on random variation alone, about 9 percent of the choices are actually inconsistent but only about 7 percent are reported as inconsistent. Thus, the reported number is roughly three-fourths of the actual number (assuming, as we have throughout, that people conform to the law of comparative judgment—especially that their dispersions are normally distributed—and also that items are listed for the DS procedure in random order). Second, regarding the specification of individual inconsistent choices, not only are some actual inconsistent choices missed by the procedure but also some of the reported inconsistent choices are not actual inconsistent choices. In fact, at $\Delta = 2$, the DS procedure correctly finds only about 53 percent of the actual inconsistent choices and 23 percent of the reported inconsistent choices are not actually inconsistent.

Despite these inaccuracies, the DS procedure can be useful in the assessment of paired comparison data. Regarding aggregate estimates, the relation between the total numbers of reported and actual inconsistent choices is essentially linear across a wide range in Δ if the items are evenly or randomly spaced along the judgment dimension of interest. Similarly, we found a nearly linear relation of total reported inconsistent choices to internal reliability as measured by ζ . These findings suggest that the DS procedure can be used to estimate the actual number of inconsistent choices and compare sets of paired comparisons for relative consistency.

Regarding the specification of individual inconsistent choices, the error in specification is not evenly spread across the choices. In fact, most of the error occurs for choices between items that are relatively close in judged value. The percent of the actual inconsistent choices that are correctly found by the DS procedure is only about 50 percent at a PSD of 0 to 2, but rises quickly beyond that to 62, 73, 85, 96, and 99 percent at PSDs of 3, 4, 5, 6, and 7, respectively. Similarly, the percent of choices that are incorrectly specified as inconsistent by the procedure drops from 50 percent at PSD of 0 to 19, 9, 4, and 1 percent at PSDs of 1, 2, 3, and 4, respectively. In other words, among choices for which $\text{PSD} \geq 2$, 60 percent of all inconsistent choices are correctly found by the procedure (40 percent are not-found), and 91 percent of the reported inconsistent choices are actually inconsistent (9 percent are bad-found). Thus, the procedure is more accurate where it matters most, for those choices that more seriously conflict with the bulk of the respondent's choices.

The DS procedure allows a link from our simulations to the real-people experiments. We found that the relationships between PSD and reported inconsistent choices, and between ζ and reported inconsistent choices, of real people correspond quite well to those of the simulations based on IID normal discriminial dispersions with $\Delta = 2$. This finding suggests that respondents on average tend to behave as if their discriminial dispersions are normal with a standard deviation of about 2.

Finally, we found that using the DS procedure in a real-people experiment to select likely inconsistent choices, presenting the likely inconsistent choices again at the end of the paired comparisons, and then replacing the original choices with the choices made when the pairs were repeated, dramatically improves the overall consistency (ζ) of a respondent's choices. Thus, the DS procedure plays a useful role when it is used in a computerized presentation of paired comparisons, both in correcting outright mistakes and in helping to produce a more accurate representation of a respondent's preferences as those preferences are refined in the course of the experiment. If the objective of a paired comparison experiment is to obtain a well-considered ordering of the items being assessed, use of the DS procedure to select pairs for retesting is warranted. To be used most efficiently, pairs selected for retesting might be restricted to those with a PSD greater than some threshold, such as a PSD of 1 or 2.

Chapter 5. Consistency and Reliability

Introduction

As defined in Chapter 1, *consistency* occurs when the respondent chooses the stimulus with the greatest expected value (V) and *inconsistency* occurs when the respondent chooses otherwise. Other things being equal, inconsistency in a choice between two items increases as the distance between the V 's of the items narrows and as the discriminial dispersion variances of the items increase. Inconsistency might also occur because of random response mistakes, such as pressing the wrong key in an experiment administered by computer, or because of respondent inattention, lack of understanding of the task, or intentional misrepresentation of preferences.

The purpose of an application of the paired comparison method is generally to estimate the relative magnitudes of the V 's of the items in the set. Because the V 's are unknown, it is not possible to empirically observe when an inconsistent choice has occurred. In an empirical application, however, the double-sort (DS) procedure can be used to find so-called *reported* inconsistent choices, which are a best guess at the actual inconsistent choices. In addition, two measures of reliability are observable: *internal reliability* as measured by the coefficient of internal reliability (ζ) or the number of circular triads (C) and *test-retest reliability*.

Relationships Between Consistency and Internal Reliability

We first ask how well internal reliability, as measured by the number of circular triads (C), indicates the level of consistency. To address this question (and others), we used a simulation model (the PCS model described in Chapter 2) that makes paired comparison choices by virtual respondents, assuming that respondents conform to normal distribution theory and the law of comparative judgment. Because the V 's are known for a simulation, the actual consistency of each choice is also known. Given the specified V 's and discriminial dispersion standard deviations (Δ) of the items in the choice set, the model allows, among other things, the estimation and comparison of measures of consistency and internal reliability.

Individual Respondent Variation

Before showing the effect of alternative levels of simulation parameters (such as Δ or number of items) on the mean result across respondents, we present the results of a single simulation to show how randomness in momentary preference causes variation among respondents. Figure 5-1 compares internal reliability measured as C with consistency measured as the actual, or true, number of inconsistent choices for a simulation with 21 items given a plausible configuration of the V 's and Δ s (randomly spaced V 's, Δ from 1 to 5 randomly assigned to the items).⁴⁴

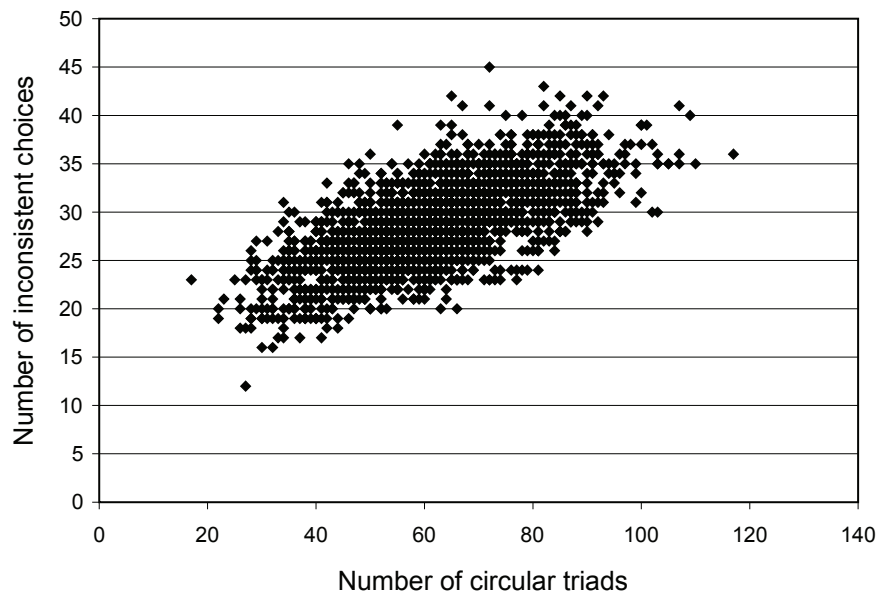


Figure 5-1. Inconsistency versus internal reliability for 2,000 virtual respondents, given 21 randomly spaced V s ranging from 1 to 21 and heterogeneous normal discriminial dispersions (Δ from 1 to 5 randomly assigned to the items).

Each data point in the figure represents one of 2,000 virtual respondents, who are identical except for momentary random variation of preference along their discriminial dispersions. The correlation (adjusted R^2) between the two measures is positive and significant but modest at 0.45. The primary reason for scatter in the relationship is that a single inconsistent choice may cause from zero to several circular triads. As demonstrated in Chapter 4, at a constant Δ , the number of circular triads caused by an inconsistent choice increases with the magnitude of the difference between the V s of the items in the pair. Another reason for the scatter is that the choice for one pair interacts with the choice for another. Some circular triads caused by an inconsistent choice for one pair might cancel circular triads that would be caused by an inconsistent choice for another pair. That is, two or more inconsistent choices might be consistent relative to each other. In sum, perfect consistency yields perfect internal reliability, but perfect reliability does not necessarily indicate perfect consistency. The number of circular triads (and thus the coefficient of internal reliability, ζ) is therefore a weak predictor of the degree of consistency at the level of an individual respondent.

Effect of Δ and Number of Items

Figure 5-2 plots the average number of inconsistent choices versus the average number of circular triads, given homogeneous discriminial dispersions at a range of Δ (and also for one set of heterogeneous Δ) holding the number of items constant at 21 with V s ranging from 1 to 21. The figure shows that both the average number of inconsistent choices and the average number of circular triads vary systematically with Δ , with the mean number of circular triads exceeding the

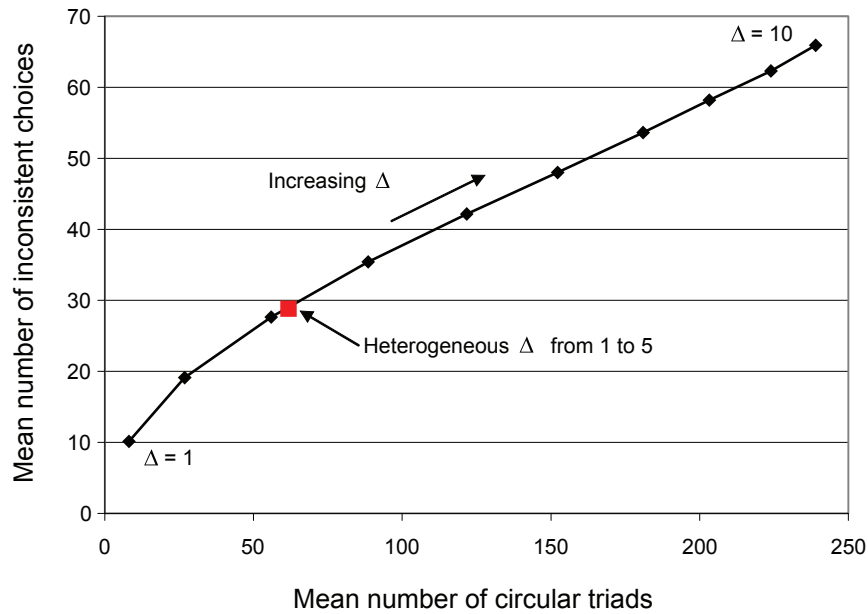


Figure 5-2. Inconsistency versus internal reliability for a variety of discriminial dispersions (IID normal discriminial dispersions with $\Delta = 1, 2, \dots, 10$, and one case of heterogeneous normal discriminial dispersions with Δ from 1 to 5 randomly assigned to the items), given 21 randomly spaced items with V s ranging from 1 to 21.

mean number of inconsistent choices for all Δ of above about 1.2. For example, at $\Delta = 2$, 19 inconsistent choices produce on average 27 circular triads, and at $\Delta = 5$, 42 inconsistent choices produce on average 122 circular triads. Beyond $\Delta = 3$, the relation is nearly linear, with each additional inconsistent choice contributing about 4.8 circular triads. However, at $\Delta = 1$, 10 inconsistent choices cause on average only 8 circular triads. At low Δ , inconsistent choices tend not to interact, thereby avoiding multiple circular triads per inconsistent choice.

Figure 5-3 shows the effect of the number of items in the choice set (t) on the relation between inconsistency and internal reliability, holding constant both Δ and the average distance between proximate V s of the items.⁴⁵ Reducing the number of items in the set reduces both the number of inconsistent choices and the number of circular triads along a nearly linear path. This finding holds for other levels of Δ . From figures 5-2 and 5-3 we conclude that the expected number of true inconsistent choices is strongly and systematically related in a given experiment to the: (1) average number of circular triads, (2) standard deviation of the discriminial dispersions of the items in the set, and (3) number of items in the set.

We combined the results of numerous simulations that differed in Δ (given IID normal discriminial dispersions) and number of items in the choice set (holding constant the average distance between proximate V s, thus changing the range in V to correspond to the number of items), and then regressed number of inconsistent choices (IC) on number of items in the set (t), Δ , and the number of circular triads (C), producing the following result, with $R^2 = 0.9994$:

$$IC = 0.76 t^{0.53} \Delta^{0.21} C^{0.43} \quad (25)$$

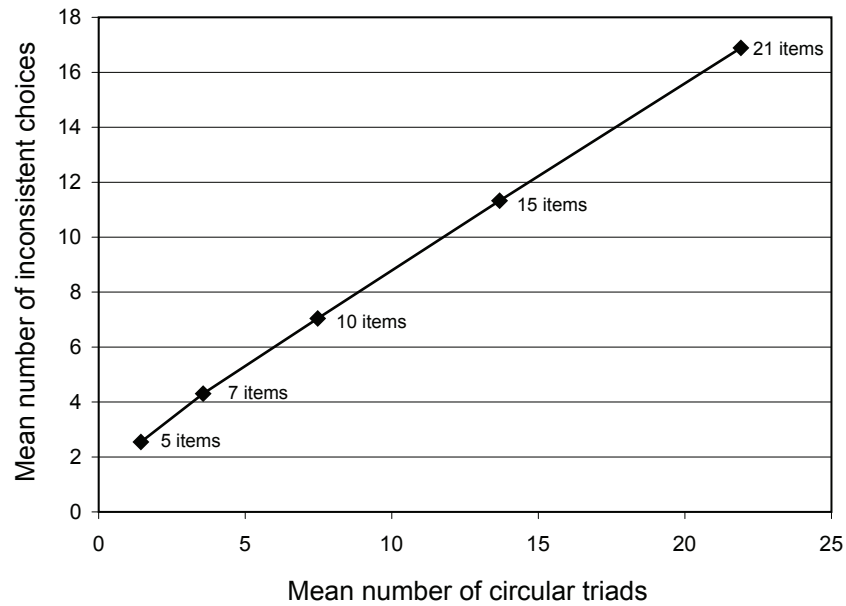


Figure 5-3. Inconsistency versus internal reliability by number of items in the choice set given evenly spaced V s and IID normal discriminial dispersions with $\Delta = 2$.

This result suggests the following four hypotheses. First, the functional form of the relationship between the expected number of inconsistent choices and the expected number of circular triads remains constant as the number of items changes. Second, at a given Δ and number of circular triads, the number of inconsistent choices increases at a slightly decreasing rate with the number of items. Third, at a given number of items and Δ , the number of inconsistent choices increases at a decreasing rate with the number of circular triads. Fourth, one empirical non-linear function approximately describes the relationship between the expected number of inconsistent choices and the expected number of circular triads. Given any empirical experiment with evenly spaced expected values and IID normal discriminial dispersions, these hypotheses suggest that the expected number of inconsistent choices is strongly predicted by the number of circular triads for any reasonable levels of t and Δ .

Figure 5-4 again compares consistency with internal reliability, but this time with reliability expressed in terms of ζ rather than in terms of C . Expressing the relation in this way enhances comparability across simulations of different numbers of items. The figure shows that varying the number of items in the set, while keeping the mean distance between proximate V s constant (thus reducing the range in V as t decreases), modifies the relationship between consistency and reliability. We see that as the number of items increases, the slope of the relationship between mean proportion consistent and mean ζ tends to decline at a decreasing rate. We also note that the curves tend to converge as the number of items increases and as ζ increases. This convergence is most obvious when mean proportion consistent and mean ζ are both about 0.90. Interestingly, in paired comparison experiments with real people (1,230 people making 170,546 choices), where the choice sets include goods and sums of money, the mean ζ was 0.93 (the median was 0.94).⁴⁶

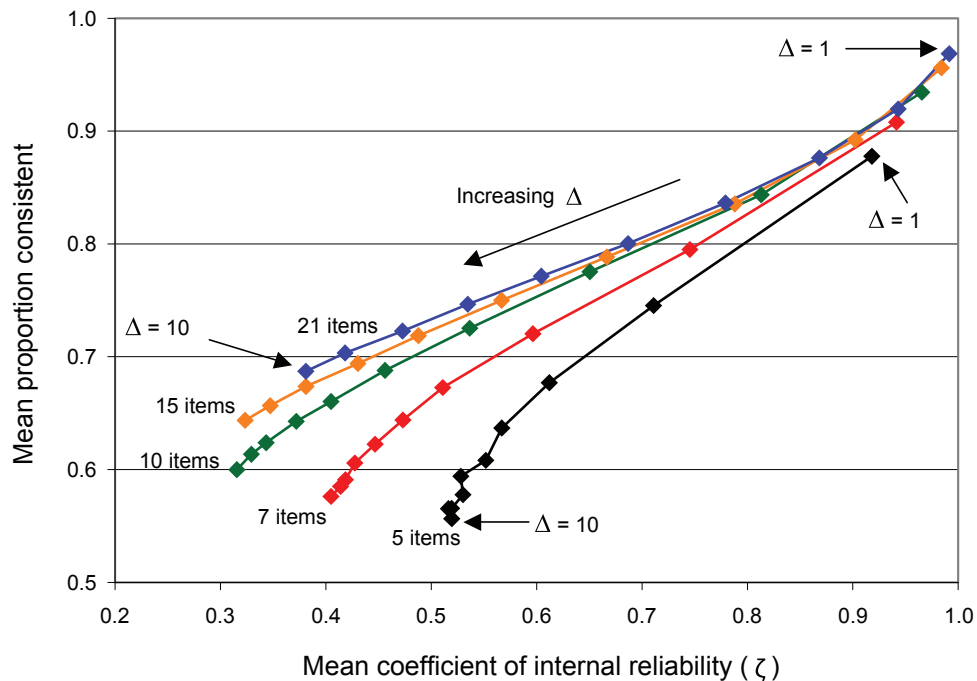


Figure 5-4. Inconsistency versus internal reliability by number of items in the choice set given randomly spaced V s, IID normal discriminial dispersions with Δ varying from 1 to 10, and holding constant the average distance between the V s (thus changing the range in V to correspond to the number of items).

Thus, assuming people in real paired comparison experiments act according to the theoretical assumptions underlying the simulations, we would conclude that the mean proportion of choices that were consistent was also about 0.93.

As previously mentioned, figure 5-2 shows how the relation of the mean number of inconsistent choices to the mean number of circular triads varies with Δ for IID normal discriminial dispersions from 1 to 10. Also depicted in the figure are the results for heterogeneous discriminial dispersions, with Δ randomly varied from 1 to 5 inclusive among the items. As seen, the result for these heterogeneous discriminial dispersions falls very near the relationship for a homogeneous dispersion of $\Delta = 3$. We hypothesize that heterogeneity of discriminial dispersion variance does not disrupt the empirical relationship found for homogeneity of variance. If the V s of the items in the choice set are randomly or uniformly distributed on the preference continuum, the expected number of circular triads strongly predicts the expected number of inconsistent choices.

Effect of Range of V

Figure 5-5 shows the effect of range of V on the relation of consistency to internal reliability, in this case for simulations with 10 items. Two ranges of V are shown, 1 to 10 and 1 to 19. The figure shows that increasing the range of V increases the intervals between the V s relative to the size of Δ , thereby reducing the overlap among the discriminial dispersions. Increasing the range of V simply reduces the number of inconsistent choices and circular triads while preserving the functional form of the relationship between the two.

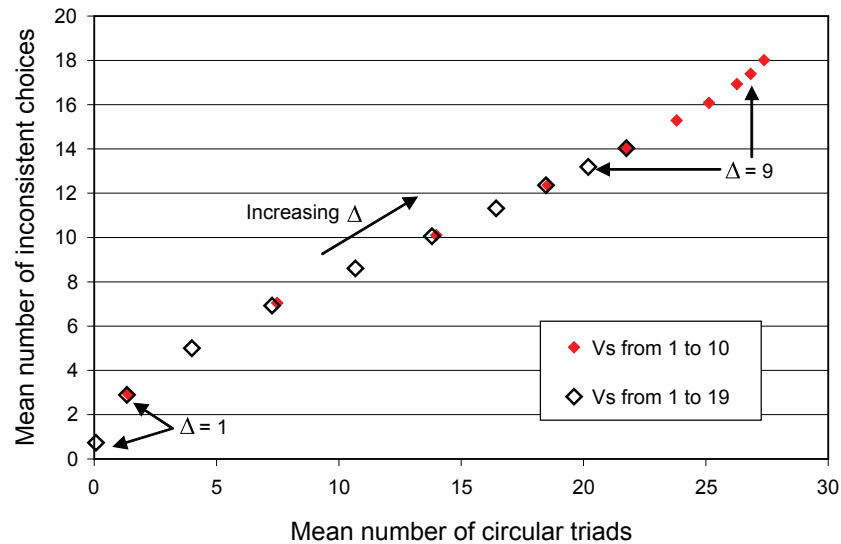


Figure 5-5. Inconsistency versus internal reliability by Δ for two ranges of V given 10 evenly spaced V s and IID normal discriminial dispersions with $\Delta = 1, 2, \dots, 10$.

Effect of Clustered Vs

The relationships described above tend to fall apart when the V s are not uniformly or randomly distributed. Figure 5-6 shows what happens when the V s are tightly clustered at opposite ends of the value continuum such that the between cluster differences in V are much greater than the within cluster differences. The

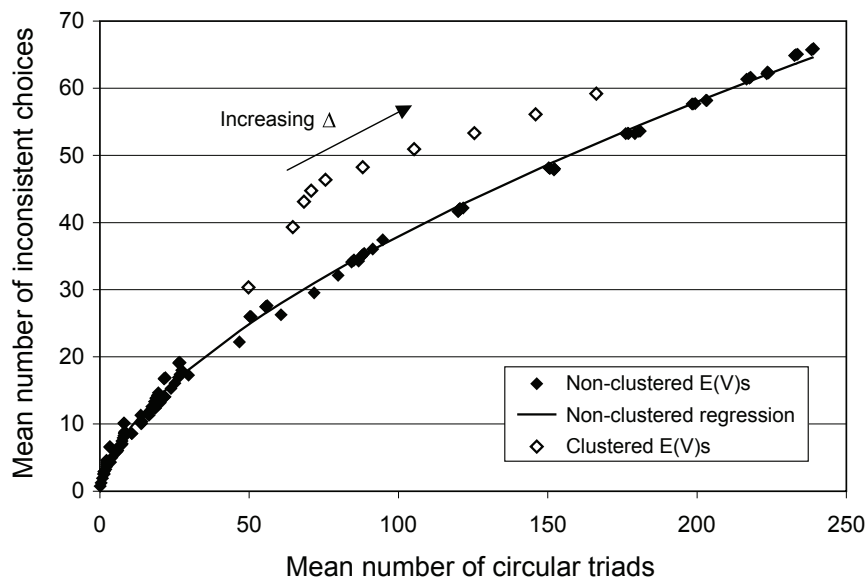


Figure 5-6. Inconsistency versus internal reliability for non-clustered and clustered V s. The points for non-clustered V s represent 120 simulations that differ by spacing of the V s (uniform or random spacing), discriminial dispersions (IID normal with $\Delta = 1, 2, \dots, 10$, or heterogeneous normal with Δ from 1 to 5 randomly assigned to the items), number of items (from 5 to 21), and range of V s (from 5 to 21). The points for clustered V s represent simulations with 21 items separated into two widely spaced clusters. The 10 simulations correspond to IID normal dispersions with a $\Delta = 1, 2, \dots, 10$.

figure shows the relation of the mean number of inconsistent choices to the mean number of circular triads. For non-clustered V s, a large mixture of simulations with uniformly or randomly spaced V s and homogeneous and heterogeneous discriminial dispersions are represented. We see that the relationship for non-clustered V s tends to follow a non-linear empirical relationship without regard to the number of items or whether the discriminial dispersion variance is homogeneous or heterogeneous. The mean number of circular triads strongly predicts the mean number of inconsistent choices (R^2 for the non-linear regression of non-clustered points = 0.9994). If the expected values are tightly clustered, however, the relationship falls apart, both for two clusters at opposite ends of the continuum (as shown in the figure) and for the other case we investigated (three clusters evenly spaced across the continuum). We conclude that to predict the expected number of inconsistent choices in a paired comparison experiment, one must design the experiment such that the V s of the items tend to be randomly or uniformly distributed across the continuum. This problem suggests a two-stage design. The first stage, which functions like a pilot study, identifies items that are clustered together and can perhaps be represented by a single item.

Test-Retest Reliability

As described in Chapter 2, the computer program that administers the paired comparison experiment to real-people retests (in other words, presents a second time) all choices identified by the DS procedure as inconsistent. The program also retests 10 randomly selected choices identified as consistent. The retests occur immediately after all possible pairs of items have been initially presented, with no perceivable break in the presentation such that respondents are most likely not aware that the retesting is occurring. These retests allow a measure of test-retest reliability. The purpose of this section is to examine test-retest reliability and the implications of the retest results for consistency.

Table 5-1 reports the aggregate results of six independent paired comparison experiments with real people (the individual experiments are described in more detail in Chapter 2—see especially table 2-1). About 7 percent (11,706) of the total of 170,546 choices were isolated by the DS procedure as inconsistent. Fifty-nine percent of those choices were switched when retested, as were 8 percent of the

Table 5-1. Reported inconsistency and choice switching in real experiments.

Study	Reported inconsistent		Inconsistent and switched		Sampled from reported consistent		Coefficient of internal reliability	
	Number	Percent	Number	Percent	Number	Percent switched	Mean	St. Dev.
Original	3,688	7.3	2256	61.2	3270	8.9	0.91	0.077
Clarke	3,830	7.1	2182	57.0	4630	8.2	0.94	0.046
Birjulin A	940	6.4	585	62.2	950	8.1	0.92	0.083
Birjulin B	1,054	7.2	613	58.2	940	9.5	0.91	0.078
Bid range 700	1,135	6.3	648	57.1	1250	6.2	0.94	0.069
Bid range 9000	1,059	5.8	624	58.9	1260	7.1	0.94	0.048
All	11,706	6.9	6908	59.0	12300	8.1	0.93	0.068

consistent choices that were retested. Respondents are obviously more likely to switch reported inconsistent choices than reported consistent choices (if this were not the case, we would expect the two percentages to be equal), suggesting that the DS procedure provides useful information about which choices are likely to be switched on retest.

As explained in Chapter 4, the likelihood that a reported inconsistent choice is a true inconsistent choice increases with PSD. At the extreme, $PSD = 0$, half of the reported inconsistent choices are in fact consistent (fig. 4-9), and on average we would expect only about half of these reported inconsistent choices to be switched on retest. As PSD increases, the accuracy of the DS procedure improves, as does the likelihood that a reported inconsistent choice will be switched on retest.

Figure 5-7 compares the switching behavior of real people with that of the simulated respondents. The proportions of pairs that are switched when retested are shown as a function of PSD, for choices reported as consistent and as inconsistent. The behavior of the real respondents is generally similar to that of the virtual respondents. In both cases, the proportion of inconsistent choices switched approaches 1 as PSD increases, and the proportion of consistent choices switched approaches zero. As expected from normal theory, the simulated proportion switched approaches 0.5 as PSD approaches zero, both for consistent and for inconsistent choices. However, for the real experiments, as PSD approaches zero, the proportions shift downward relative to those of the simulations, with the proportion of reported inconsistent choices switched approaching 0.43 and the proportion of reported consistent choices switched approaching 0.36.

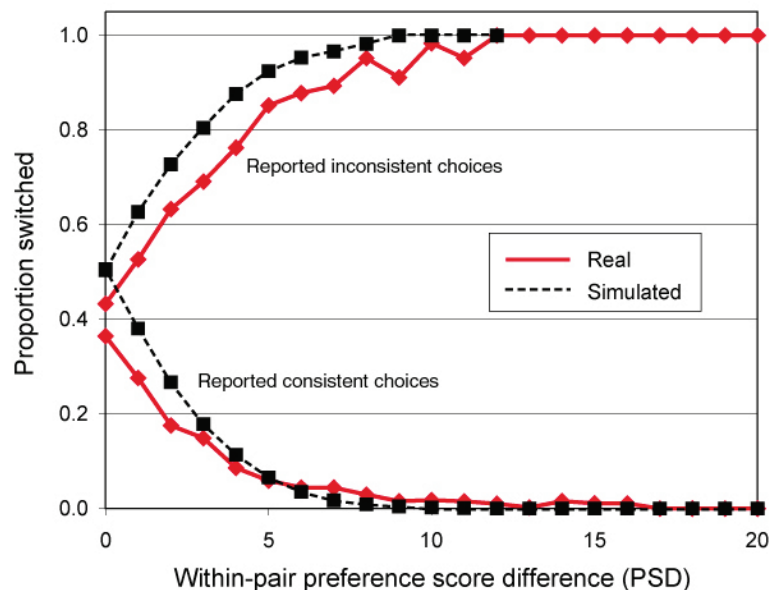


Figure 5-7. Theoretical versus actual choice switching for choices reported as consistent and inconsistent by the DS procedure. The real-people results are based on 170,546 original choices. The simulation results are based on 1,260,000 original choices for evenly spaced Vs and 1,260,000 original choices for randomly spaced Vs, given 21 items, IID normal discriminial dispersions with $\Delta = 2$, and a random input order of the items in the raw choice matrix.

The results from the real experiments raise two interesting questions. First, why do the real study proportions shift downward as PSD approaches zero? According to the normal theory, choices should become increasingly random as PSD approaches zero. Second, why do the proportions for consistent and inconsistent choices not converge to the same proportion at $\text{PSD} = 0$? Normal theory would again imply that at $\text{PSD} = 0$ choices should be purely random. Those questions require further study and we can only speculate here.

Regarding the first question, it must be noted that a PSD of zero does not necessarily mean that the two items in the pair have exactly the same V in the mind of the respondent or that there is no tendency to prefer one to the other. A PSD of zero means that the two items were each preferred over the same number of other items in the choice set. In fact, in the choice involving only the two items, one might have been consistently chosen over the other because of a real perceived value difference. If so, there would be a tendency for reported consistent choices to resist switching when retested even when $\text{PSD} = 0$.

Another possibility for a resistance to switching is the application of other than a compensatory decision rule. One hypothesis for why people sometimes make the same choice upon retesting, even when the two items are perceived as being of nearly the same value, is that people sometimes remember their prior choices. Such a memory effect would, if it exists, be most apparent at $\text{PSD} = 0$ and would tend to disappear as the difference increases, for at larger PSDs preferences are clearer and memory of prior choices becomes irrelevant. A second closely related hypothesis is that people tend to develop simplifying rules for making difficult choices and those rules tend to remain in effect for all choices except those for which there are overwhelming reasons to ignore the rule. One example of such a rule would be that people tend to cue on specific attributes in specific pairs. For example, a person might tend to choose a public good over a private good whenever the values of the two goods are otherwise similar.

The second question is more perplexing, and it cannot be ignored because the two proportions, 0.43 and 0.36, differ significantly. In the first case, 1,346 of 3,110 choices were reversed. In the second, 86 of 236 choices were reversed. A normal test for the difference between two proportions yields a standard normal deviate of 2.10 and a two-tailed probability of 0.036, which is of course significant at the 0.05 level. The two-tailed probability assumes no a-priori reason to expect one proportion to be larger than the other. A one-tailed test assumes that we have reason to expect reported inconsistent choices to be more likely switched than reported consistent choices and yields a probability of 0.018. The answer to this perplexing finding must await further research.

Evidence of Learning and/or Memory

If the theoretical assumptions of the law of comparative judgment hold in paired comparison experiments, choices will be independent. Under full independence, the likelihood that a choice will be inconsistent or reversed when retested will be independent of the sequence in which the pairs are presented. We examined the data from our real-people experiments to see if subjects' behavior appeared to satisfy the independence assumption.

For the pooled results of all six real-people experiments, the proportion of choices reported as inconsistent is strongly dependent on sequence of choice over the first 35 choices (fig. 5-8). The slope of proportion by sequence is significantly negative ($p = <0.0001$). For choices in the sequence range of 36 and higher, the slope is also negative but not significant ($p = 0.059$).

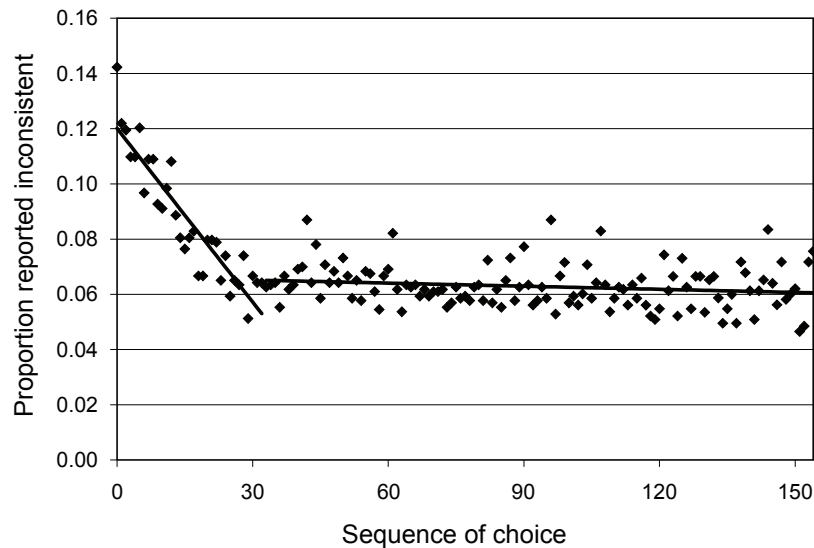


Figure 5-8. Proportion of choices reported as inconsistent by the DS procedure versus sequence of the choice for a composite of the six real-people experiments.

Figure 5-9 shows, for the composite of the real-people experiments, the relationship between the proportions of choices switched on retest and the sequence of choice, both for choices reported as consistent and inconsistent. In both cases, the proportion switched declines significantly as sequence increases, with the most rapid decline occurring over the first 10 choices. Although the average slope,

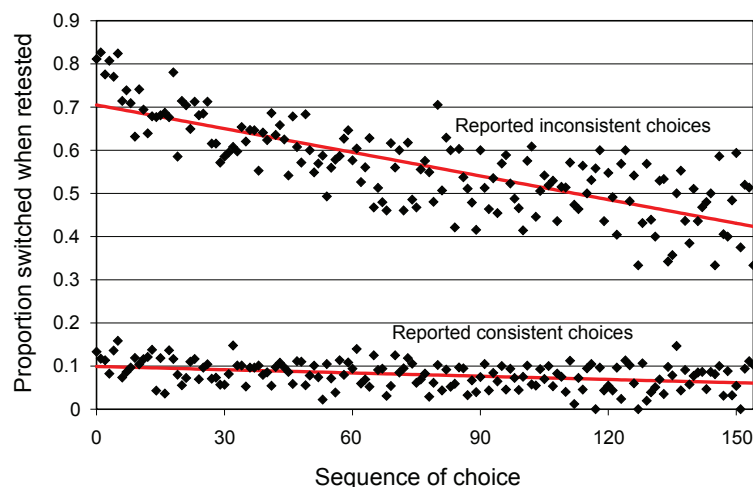


Figure 5-9. Proportion of choices that were switched when retested versus sequence of the choice for a composite of the six real-people experiments.

as estimated by linear regression, is much steeper for inconsistent choices than for consistent choices, the total percent change in overall linear slope (straight lines in fig. 5-9) is nearly the same for the two sets of choices (39.93 percent for inconsistent choices and 39.15 percent for consistent choices). In both cases, the downward slope is statistically significant. It is notable that nearly 80 percent of the first six choices encountered and reported as inconsistent were switched when retested, whereas only 47 percent of the last six choices encountered and reported as inconsistent were switched when retested.

Figure 5-8 shows that inconsistency persists even after 100 or more choices, and figure 5-9 shows that some switching occurs even for choices that were encountered only a few pairs before the retests began. This persistence of inconsistency and switching suggests that many of the reported inconsistent choices are random (which can occur when they are associated with small PSDs) and that some of the reported inconsistent choices probably are mistakes that are avoided when those choices are retested. However, the two figures also show conclusively that choices are not independent of sequence. If the respondents behaved strictly according to the theory of the law of comparative judgment, the proportion inconsistent and the proportion switched would be independent of sequence. If later choices are somewhat dependent on prior choices, two possibilities arise. First, the respondents become more confident in their preferences as they become more familiar with the items in the experiment (essentially, respondents' Δ s become smaller) and second, the respondents learn, construct, or discover their values as the experiment progresses (resulting in respondents' V 's changing over the course of the experiment). Of course, both of these behaviors might occur at once.

The sample sizes of the experiments are insufficient for computing average preference scores for each item for each choice in the sequence of choices, but we can compare average preference scores from sets of choices, such as sets of early versus late choices. For a given experiment and set of sequential choices, the average preference score of each item was computed as the number of times, across all respondents in the experiment, the item was chosen over the number of times the item appeared among the choices. Such computations are possible because of the random ordering of pairs for each respondent. A comparison of the average preference scores estimated from the first 30 choices with those estimated from choices 71 to 100 produced correlations of early versus late average preference scores ranging from 0.96 to 0.99 across the experiments. This indicates very little shift in preferences with sequence, thus supporting the claim that only the disturbance terms were changing over the course of the sessions.

A behavior that could be associated with the hypothesis of diminishing dispersion with sequence is that respondents gradually develop and apply simplifying rules that persist from choice to choice. An additional hypothesis that applies to retests is that respondents tend to remember previous choices—perhaps to avoid appearing to be inconsistent or simply by subliminal memory—and that memory is more effective the more recently the prior choice was made. We suspect that both of these behaviors are in play among our actual respondents and that the relative importance of the behaviors varies across respondents. Unfortunately, our data do not allow us to test for these behaviors.

Still another indication of increasing familiarity with the items is the change over sequence in the amount of time respondents take to make their choices. Figure 5-10 shows the relation of mean decision time to sequence of choice pooled over the six real-people studies. Mean decision time starts out quite high and declines rapidly as respondents familiarize themselves with the experiment and the items among which they are making choices.⁴⁷ Further, this drop in decision time occurs for both consistent and inconsistent choices (fig. 5-11).⁴⁸ Thus, even for consistent choices, some amount of familiarization is apparently occurring.

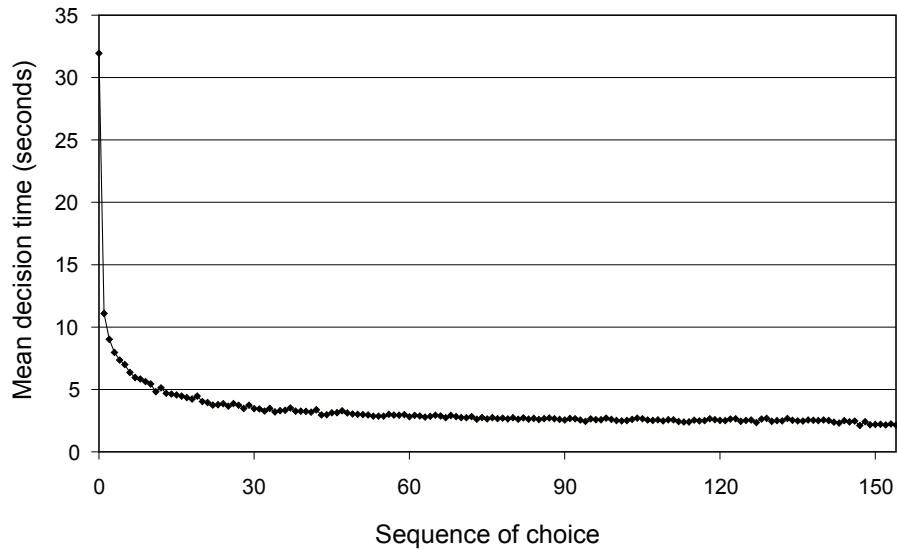


Figure 5-10. Decision time versus sequence for a composite of the six real-people experiments.

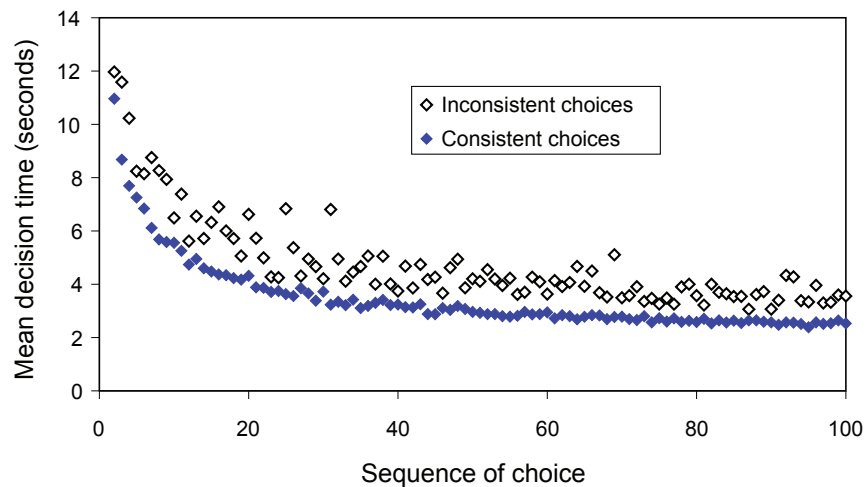


Figure 5-11. Decision time versus sequence for a composite of the six real-people experiments, showing the 2nd through the 100th choice for both consistent and inconsistent choices.

Does Choice Difficulty Influence Consistency and Choice Switching?

Consistency in paired comparisons is, we suspect, largely a matter of (1) respondent ability and effort and (2) the difficulty of the choices. The difficulty of a choice is in turn a matter of how close the items are on the dimension of interest (close calls being more difficult than obvious ones) and of the complexity of the items. PSD is a measure of how close the items are on the dimension of interest. Complexity may be related to the nature of the items being compared. For example, a public good is probably more complex than a commonly available private good, which in turn is probably more complex than a monetary amount. Thus, if the time taken to make a choice is an indication of choice difficulty, time to make a choice should also be related to PSD and the nature of the items compared in the choice. Similarly, if inconsistency is an indication of difficulty, we would expect to find that respondents take longer to make inconsistent choices than to make consistent choices, and that inconsistency is more likely for more complex choices.

Decision Time

Figure 5-12 shows the relation of mean decision time to PSD pooled over all six real-people studies. Decision time increases markedly as PSD declines, indicating that choices between items of similar value are more difficult to make.

Figure 5-11 shows that, for each of the first 100 pairs encountered, reported inconsistent choices took more time than reported consistent choices. Figure 5-13 further examines this relationship, showing a strong tendency for decision time to increase with the likelihood that a choice will be reported as inconsistent for decision times of 20 seconds or less. At decision times above 20 seconds, the points become very scattered because of declining sample size and thus are not shown in the figure. The figure suggests that more difficult choices take longer and are more likely to be inconsistent with the respondent's other choices.

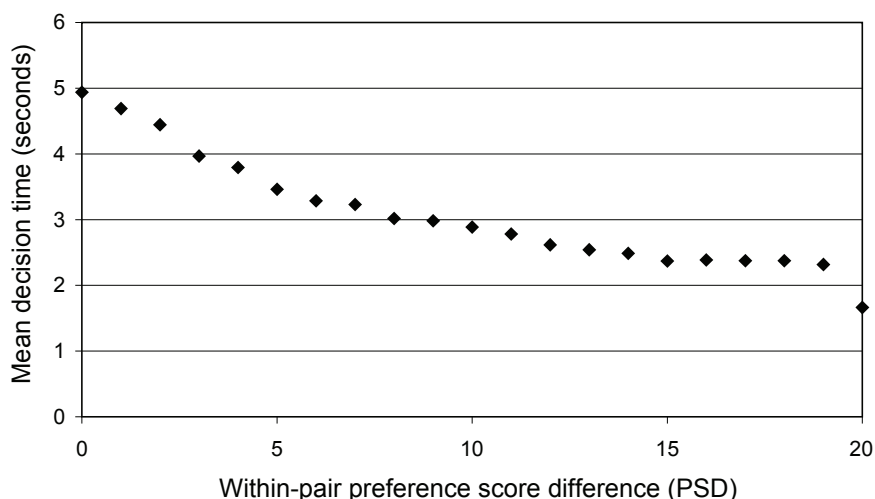


Figure 5-12. Decision time versus PSD for a composite of the six real-people experiments.

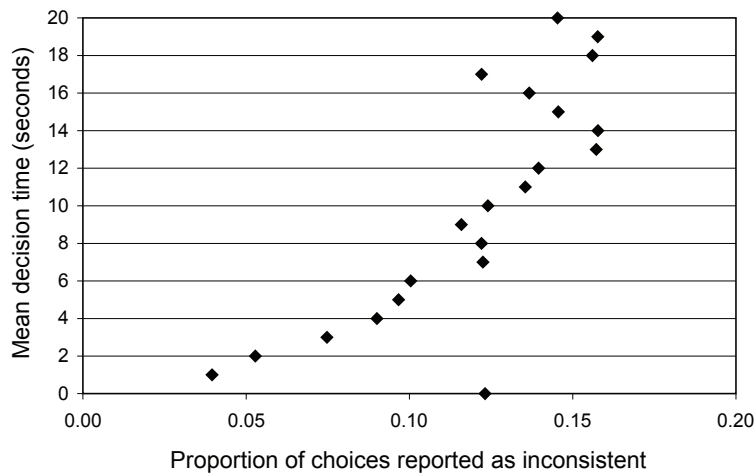


Figure 5-13. Decision time versus reported inconsistency (based on the DS procedure) for choices taking 20 seconds or less for a composite of the six real-people experiments.

Type of Item

We now ask whether choice difficulty varies with the type of items involved in the choice. One measure of choice difficulty is the proportion of choices that were reported as inconsistent by the DS procedure. As seen in table 5-2, the percent of the choices that were reported inconsistent is lowest for choices involving a monetary amount and highest for choices involving two goods where at least one of the goods is a public good. The greater inconsistency for public good choices than for private good choices holds for choices involving money (the 6.6 percent is significantly greater than the 5.5 percent in table 5-2) but not for choices involving only goods (the 8.2 percent is not significantly greater than the 7.7 percent).

Table 5-2. Mean decision time and occurrence of inconsistency by type of choice, for pooled data.

Type of choice	Number of choices	Percent inconsistent	Mean decision time (seconds)
Private good vs. money	51,922	5.5	3.09
Public good vs. money	67,441	6.6	3.17
Private good vs. private good	8,384	7.7	3.97
Public good vs. public good	14,880	8.2	4.53
Private good vs. public good	27,919	8.9	4.23
All choices	170,546	6.9	3.48

Choices requiring comparisons across types of goods showed significantly greater reported inconsistency than did choices between private goods (the 8.9 percent is significantly greater than the 7.7 percent), but not greater reported inconsistency than choices between public goods (the 8.9 percent is not significantly greater than the 8.2 percent).

Table 5-2 also reports the mean decision times for the five combinations of items, showing that decisions between goods and sums of money require less time than decisions between goods, and that decisions between goods, where one or both are public goods, tend to require the most time. The Scheffé multiple comparison test shows that each of the five measures of mean decision time in the table is significantly different from each of the other four (at the 0.95 probability level) except for the two mean decision times for choices involving monetary amounts (3.09 and 3.17 seconds).

Effect of Retest-Corrected Choices on Internal Reliability

As seen in figure 4-15, reversing (in other words, correcting) choices that were switched when retested substantially increases respondents' coefficient of internal reliability (ζ). Figure 5-14 compares the raw and corrected ζ profiles from the respondents of the six real-people experiments, as presented in figure 4-15, with similar profiles from simulations with randomly spaced V 's and discriminial dispersion Δ s of 1, 2, and 3. Two conclusions are suggested by this comparison.



Figure 5-14. Retest-corrected versus uncorrected coefficient of internal reliability (ζ) for a composite of six real-people experiments compared to simulations given 21 items with randomly spaced V 's, IID normal discriminial dispersions with Δ of 1, 2, or 3, and a random raw matrix input order for the items. Simulated curves are shown with black lines. The lightly colored lines indicate real-people results.

First, real people on average demonstrate a Δ at the low end of the range we have examined using the simulation model (Δ s of from 1 to 10). For uncorrected results, the internal reliability of the real respondents is, on average, close to that of the simulation with $\Delta = 2$ (mean ζ s are 0.928 and 0.931, respectively). And for corrected results, the internal reliability of the real respondents is very similar to that of the simulation with $\Delta = 1$ (mean ζ s are 0.980 and 0.979, respectively).

Second, the real respondent profile does not have the characteristic profile produced by the IID normal dispersion assumption on which the simulations are based. The simulation profiles initially (at the left) are strongly concave upward. They inflect at the median and then are strongly concave downward at the other extreme. Those characteristics are the expected outcome of a homogeneous population of respondents with IID normal discriminial dispersions. The same characteristics would be typical of any symmetrical distribution with a strong central tendency, as with the distributional assumptions that underlie modern logit- or probit-based discrete choice random utility models.

The real respondent profile is initially concave upward, but the curvature is much weaker than that of the simulation profiles and is barely discernable. At the other extreme, the concave downward curvature of the real respondent profiles is much stronger than that of the simulations. As seen by comparing the uncorrected real respondent curve with the simulated curves, about two-thirds of the real respondents have a ζ indicative of a Δ less than 2 and a few (less than 5 percent) have a ζ indicative of a Δ greater than 3. Further, a small minority of respondents have a ζ indicative of a Δ very much below 2. We conclude that (1) the real respondent sample is very heterogeneous, (2) a substantial proportion of the respondents exhibit a notable degree of consistency, and (3) the sample is strongly skewed toward the inconsistent end of the profile. Far more respondents are at the inconsistent end of the profile than would be expected by projecting the curvature from the consistent end. This finding suggests that a few respondents are responding abnormally to the experiment, raising the possibility that those with a ζ less than some threshold magnitude might reasonably be dropped from the sample. However, any such threshold criterion would be arbitrary.⁴⁹

These profiles suggest that samples of real people are likely to be very heterogeneous and that the heterogeneity is likely to be strongly skewed toward the less consistent end of the spectrum. Such heterogeneity strongly violates the assumptions that underlie modern discrete choice theory and random utility models. It is obvious that we cannot assume that the combined effect of intra-personal and inter-personal dispersion emulates IID intra-personal dispersion of a homogeneous sample of respondents as is apparently assumed in empirical applications of random utility theory. It would be useful in an actual application to (1) screen respondents by their ability to respond meaningfully and (2) segment the sample so as to reduce heterogeneity.

Summary

Assuming that people conform to the law of comparative judgment, the simulations have allowed us to determine how accurately our measures of internal reliability (C and ζ) predict the actual number of inconsistent choices. At the level

of the individual respondent, C is only roughly related to the actual number of inconsistent choices, but on average across a group of respondents, C and ζ vary systematically and nearly linearly with the true number of inconsistent choices as long as the items are spread more or less evenly along the V continuum. But can we accept the assumption of IID normal discriminial dispersions that underlies the simulations? The evidence is mixed. Evidence presented in Chapter 4 suggests that groups of real people correspond on *average* quite well to those of the simulations based on IID normal discriminial dispersions of $\Delta = 2$. Further, evidence from the current chapter suggests that, given a realistic number of items to be assessed (say 10 to 20) and discriminial dispersions of $\Delta = 2$, the *mean* ζ is about 0.9. Notably, the mean ζ across our 1,230 actual respondents, uncorrected, is very similar at 0.93. However, evidence from the current chapter also suggests that those means reflect considerable heterogeneity across the respondents in both ζ and Δ .

Observation that choice consistency improves over the course of the paired comparison experiment and correction of choices based on retesting of potentially inconsistent choices tends to improve respondents' overall measures of internal reliability offers strong evidence of a process of gradual familiarization with the items being assessed, with associated narrowing of the discriminial dispersions. This evidence suggests that the independence assumption is open to question—later choices do appear to depend on earlier ones.

The assumption of identical disturbance distributions also appears to be violated, as suggested by the finding that people differ substantially in internal reliability (ζ). Further, when measures of ζ for actual people, without correction, are compared with those for the simulations, we find that most people's choices suggest a mean $\Delta < 2$, but other peoples' choices suggest a mean Δ above 2 or even 3. Also, a small minority appears to be so unreliable in their choices as to suggest that they either did not understand what they were asked to do or made no attempt to answer carefully or honestly. A screening for this small minority seems in order.

Chapter 6. Closing Thoughts

Our data, both from the actual paired comparison experiments and the simulations, clearly show that most inconsistencies occur among choices that are close calls (in other words, of small PSD) where, based on Thurstone's law of comparative judgment, we would expect inconsistencies to occur. Further, we found that inconsistencies that occur that are clearly not close calls (those of large PSD) tend to be corrected upon retesting, suggesting that the original choices were mistakes or at least poorly considered choices.

The lack of precision in binary choice that is characterized by the law of comparative judgment and evident in our data from several paired comparison experiments is not necessarily an indication of a lack of rationality on the part of respondents. People have perceptual and cognitive limitations that in turn limit their ability to consistently make binary choices, especially among items of similar value. For this reason, choice must be viewed as a stochastic process and transitivity must be viewed as a stochastic concept. On this point there is a considerable literature and substantial agreement.

We should mention one additional potential source of random inconsistency in our paired comparison data: a lack of consequences of a poor choice due to the hypothetical nature of the paired comparison experiments. Lacking a payoff for careful and honest revelation of preferences, some respondents may not fully apply themselves in making their choices. However, our data suggest that most respondents do their best and that only a small minority of respondents fail to take a paired comparison exercise more or less seriously. Comparisons with choices made in settings with meaningful consequences would help determine the level of hypothetical bias associated with paired comparison experiments.

Because the double-sort (DS) procedure relies on a respondent's preference scores as the measure of the true (expected) values of the items, the procedure does not always accurately select individual inconsistent choices (especially for pairs of low PSD) and underestimates the total number of actual inconsistent choices, assuming normal discriminial dispersions. However, the simulations show that, despite the inaccuracies, the procedure is unbiased and thus can be used to (1) compare sets of choices for relative consistency and (2) usefully select individual pairs for retesting, at least for pairs with a PSD above some threshold. First, using the procedure to compare sets of choices, we found that consistency was greater for some types of items than for others. For example, choices including a monetary amount as one of the items were more consistent than those between two goods, and choices involving a commonly available private good were more consistent than choices involving a public good. Inconsistent choices are less likely the larger is PSD. As PSD approaches zero, choices become harder for respondents to make. This is reflected in inconsistent choices having longer response times than consistent choices. Second, using the procedure to select pairs for retesting and then replacing the original choices with the retest choices led to a significant improvement in the internal reliability of respondents' choices. Thus, if the objective of a paired comparison exercise is to efficiently obtain a well-considered ordering of the items in

the choice set, use of the DS procedure along with retesting of inconsistent choices is warranted.

Respondents in our experiments acted on average as if their discriminial dispersions had a standard deviation (Δ) of about 2 given choice sets of about 20 items ranging in V from 1 to 21, corresponding to a coefficient of internal reliability (ζ) of about 0.93. Furthermore, when respondents' original choices were corrected based on the results of retesting of choices isolated by the DS procedure as inconsistent, the respondents' apparent average Δ was reduced to about 1, corresponding to a ζ of about 0.97. However, these averages do not tell the whole story. Respondents differed more in internal reliability, one from the next, than can be explained by random variation alone at a constant Δ , leading us to conclude that respondents vary in mean Δ . Indeed, we found a continuum, from a few who demonstrate remarkable consistency to a few who are so inconsistent that they appear to have made no attempt to respond carefully and honestly. Also, as mentioned, we found evidence that Δ varies across types of items. Respondents thus did not exhibit identically distributed discriminial dispersions, either across items or across respondents, as would be assumed for application of scaling procedures commonly used with binary choice data.

Use of the DS procedure permitted another interesting comparison of consistency among different groups of choices, in this case, of early choices with later choices. We found a dramatic increase in consistency over the first 30 or so choices, followed by a much more gradual increase over the remaining choices. As respondents became more familiar with the items being compared, they became more consistent in their choices. This improving consistency is unexpected, and it raises at least two important issues. First, the improving consistency indicates a violation of the independence assumption. Discriminal dispersions across a set of paired comparisons are apparently not independent. This order effect raises further questions about the analysis procedures, which commonly rely on the IID assumption.

Second, we must ask why this improvement in consistency occurs. Memory, of course, is one possibility, but we seriously doubt that many subjects were able to remember many of their prior choices. We find that when faced with a large number of choices, it is very difficult to keep track of all but the most recent ones (although people of extraordinary memory may rely on that ability). Another possibility is that people gradually, over the course of considering the various pairs, develop simplifying rules—such as the rule to always select the least costly item or to always prefer a private good over a public good—that they then apply when responding to any remaining pairs for which the choice is not otherwise obvious. To the extent that such rules contradict more well-considered, independent choices, they interfere with the goal of the paired comparison exercise. A third possibility, not to be discounted, is that respondents simply firm up their innate preferences as they proceed through the pairs. If memory is the explanation, the improving consistency as respondents proceed through the pairs is of little interest. If simplifying rules are the explanation, and if application of the rules minimizes respondent effort at the expense of the revelation of preferences, one must wonder about the legitimacy (and the context dependence) of the later choices. And, if a firming up of one's innate preferences is the explanation, the use of a multi-response approach like paired comparison is vindicated, and one must wonder about the

efficacy of relying on the very first response or on a method that requires only a single response. Determining the relative importance of these different possible explanations for the change in inconsistency with sequence is an intriguing and challenging research topic.

Regarding scaling of binary choice data, we restricted our analysis to methods that had been specifically developed to analyze paired comparisons, and thus did not examine modern discrete choice analysis procedures. We found that in most cases mean preference scores are about as effective as more complex paired comparison scaling methods and far simpler to calculate and use. We also found that the assumption of identically distributed error terms can be violated without serious consequences. However, we did unearth a troubling caveat: clustering of the expected values (V s) of the items along the evaluative dimension interferes with the ability of all the paired comparison scaling procedures, although in certain cases (for example, unevenly spaced clusters of V s) the psychometric scaling procedures can offer a substantial improvement over mean preference scores. Especially given small discriminial dispersions, clustering of V s concentrates the overlap among dispersions within clusters and may leave no overlap between clusters, limiting the ability of the data to reveal relationships between clustered groups of items. In general, one should avoid clustered V s and strive for at least randomly spaced V s. Finally, however, we note that in an experiment with real people, clustering is unlikely to pose as much of a problem as it did in our simulations because the simulations assumed a homogeneous sample of respondents whereas a real sample of individuals is very unlikely to have homogeneous preferences. Heterogeneity of preferences, as well as the typical level of dispersion about the V s, is likely to ameliorate the problem of clustering. This possibility suggests that future research should simulate groups of individuals with realistically heterogeneous preferences. Further, the ability of modern discrete choice procedures to handle clustered V s should be explored.

The method of paired comparisons is a powerful tool for eliciting judgments and preferences. Because the method uses binary choices for eliciting preferences, it is easily used with diverse publics and requires only the simplest of instructions. Although easy to use, the method allows for sophisticated analysis that provides a rich understanding of individual choice behavior and a reliable measure of preference. The utility of the method is enhanced by use of the computer program described herein for presenting the binary choices, which can be implemented on individual computers or over the web. The program is available for use free of charge at the website listed in Chapter 2.

The method of paired comparisons can be used to obtain preference or judgment information regarding all manner of stimuli, including goods, services, policies, states of the environment, and management alternatives. The method is regularly used by industry to assess public preferences or management options. In the realm of natural resource management, the method could be used, for example, with members of the public who are asked for their preferences regarding alternative enhancements to a national forest or with land managers asked for their judgments about the relative promise of alternative policy options for dealing with a pressing problem such as climate change.

Literature Cited

- Ben-Akiva, Moshe; Lerman, Steven R. 1985. Discrete choice analysis: theory and application to travel demand. Cambridge, MA: Massachusetts Institute of Technology. 390 p.
- Birjulin, Andrej A. 1997. Valuation of public goods using the method of paired comparison: the effects of psychological distance between goods on consistency of preferences. Unpublished Dissertation, Colorado State University, Fort Collins, CO. 118 p.
- Bock, R. Darrell; Jones, Lyle V. 1968. The measurement and prediction of judgment and choice. San Francisco, CA: Holden-Day, Inc. 370 p.
- Bradley, Ralph Allan; Terry, Milton E. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39(3/4): 324-345.
- Brown, Thomas C.; Kingsley, David; Peterson, George L.; Flores, Nick; Clarke, Andrea; Birjulin, Andrej. 2008. Reliability of individual valuations of public and private goods: choice consistency, response time, and preference refinement. *Journal of Public Economics* 92(7): 1595-1606.
- Brown, Thomas C.; Nannini, Dawn; Gorter, Robert B.; Bell, Paul A.; Peterson, George L. 2002. Judged seriousness of environmental losses: reliability and cause of loss. *Ecological Economics* 42(3): 479-491.
- Brown, Thomas C.; Peterson, George L. 2003. Multiple good valuation. In: Champ, Patricia A.; Boyle, Kevin; Brown, Thomas C., eds. A primer on non-market valuation. Norwell, MA: Kluwer Academic Publishers: 221-258.
- Brown, Thomas C.; Peterson, George L.; Brodersen, R. Marc; Ford, Valerie; Bell, Paul A. 2005. The judged seriousness of an environmental loss is a matter of what caused it. *Journal of Environmental Psychology* 25(1): 13-21.
- Buhoff, Gregory J.; Leuschner, William A. 1978. Estimating psychological disutility from damaged forest stands. *Forest Science* 24(3): 424-432.
- Clarke, Andrea. 1996. The paired comparison technique: a subterfuge for estimating relative non-market values for the things we love? Unpublished Dissertation, Colorado State University, Fort Collins, CO. 64 p.
- Clarke, Andrea; Bell, Paul A.; Peterson, George L. 1999. The influence of attitude priming and social responsibility on the valuation of environmental public goods using paired comparisons. *Environment and Behavior* 31(6): 838-857.
- David, Herbert A. 1988. The method of paired comparisons (Second ed.). New York, NY: Oxford University Press. 188 p.
- Davidson, Mark L.; McGuire, Dennis P.; Chen, Tsuey-Hwa; Anderson, Ronald O. 1995. Testing the equality of scale values and discriminial dispersions in paired comparisons. *Applied Psychological Measurement* 19(3): 257-267.
- Davidson, Roger R.; Farquhar, Peter H. 1976. A bibliography on the method of paired comparisons. *Biometrics* 32(2): 241-252.
- Dillon, William R.; Kumar, Ajith; de Borrero, Melinda Smith. 1993. Capturing individual differences in paired comparisons: an extended btl model incorporating descriptor variables. *Journal of Marketing Research* 30: 42-51.
- Fechner, G.T. 1860. *Elemente der psychophysik*. Leipzig: Breitkopf and Hartel.
- Gregory, Robin; Lichtenstein, Sarah; Slovic, Paul. 1993. Valuing environmental resources: A constructive approach. *Journal of Risk and Uncertainty* 7: 177-197.
- Guilford, J. P. 1954. *Psychometric methods* (Second ed.). New York: McGraw-Hill. 597 p.
- Hendel, Darwin D. 1977. Intransitivity on paired-comparisons instruments: The relationship of the total circular triad score to stimulus circular triads. *Applied Psychological Measurement* 1(3): 403-411.
- Hensher, David A.; Rose, John M.; Greene, William H. 2005. *Applied choice analysis: A primer*. New York: Cambridge University Press. 713 p.
- Kendall, M. G.; Smith, B. B. 1940. On the method of paired comparisons. *Biometrika* 31: 324-345.
- Luce, R. D. 1959. *Individual choice behavior*. New York: Wiley. 153 p.
- Luce, R. Duncan; Suppes, Patrick. 1965. Preferences, utility and subjective probability. In: Luce, R. Duncan; Bush, Robert B.; Galanter, Eugene, eds. *Handbook of mathematical psychology*, Vol. 3. New York: Wiley: 249-410.
- Manski, Charles F. 1977. The structure of random utility models. *Theory and Decision* 8: 229-254.
- Maxwell, A. E. 1974. The logistic transformation in the analysis of paired comparison data. *British Journal of Mathematical and Statistical Psychology* 27: 62-71.

- McFadden, Daniel. 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P., ed. *Frontiers in econometrics*. New York: Academic Press: 105-142.
- McFadden, Daniel. 2001. Economic choices. *American Economic Review* 91(3): 351-378.
- Nannini, Dawn K. 2003. Feasibility of negotiated compensation as an economic solution to environmental injustice. Unpublished Ph.D., Colorado State University, Fort Collins, CO.
- Nunnally, Jum C. 1976. *Psychometric theory*. New York: McGraw Hill. 640 p.
- Peterson, George L.; Brown, Thomas C. 1998. Economic valuation by the method of paired comparison, with emphasis on evaluation of the transitivity axiom. *Land Economics* 74(2): 240-261.
- Plott, Charles R. 1996. Rational individual behavior in markets and social choice processes: The discovered preference hypothesis. In: Arrow, K.; Colomatto, E.; Perleman, M.; Schmidt, C., eds. *Rational foundations of economic behavior*. London: Macmillan and St. Martin's: 225-250.
- Rieskamp, Jorg; Busemeyer, Jerome R.; Mellers, Barbara A. 2006. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature* 44(3): 631-661.
- Rosenberger, Randall S.; Peterson, George L.; Clarke, Andrea; Brown, Thomas C. 2003. Measuring dispositions for lexicographic preferences of environmental goods: Integrating economics, psychology, and ethics. *Ecological Economics* 44(1): 63-76.
- Rosenberger, Randall S.; Peterson, George L.; Loomis, John B. 2002. Applying a method of paired comparisons to measure economic values for multiple good sets. *Journal of Agricultural and Applied Economics* 34(1): 215-229.
- Thurstone, L. L. 1927a. A law of comparative judgment. *Psychology Review* 34: 273-286.
- Thurstone, L. L. 1927b. The method of paired comparisons for social values. *Journal of Abnormal Social Psychology* 21: 384-400.
- Thurstone, L. L. 1927c. Psychophysical analysis. *American Journal of Psychology* 38: 368-389.
- Titchener, Edward Bradford. 1901. *Experimental psychology: A manual of laboratory practice*. Volume 1: Qualitative experiments. New York: Macmillan.
- Torgerson, Warren S. 1958. *Theory and methods of scaling*. New York, NY: John Wiley & Sons. 460 p.
- Tversky, Amos; Russo, J. Edward. 1969. Substitutability and similarity in binary choices. *Journal of Mathematical Psychology* 6: 1-12.

Appendix: RUM and the Law of Comparative Judgment

The model of human judgment proposed by Thurstone and the modern RUM model used in economics and other disciplines, although related historically (McFadden 2001) and very similar in notation, differ markedly in intent. Here we present a brief comparison of the two random utility models. For this exposition, we refer to the perspectives of economists and psychologists in a stereotypical way, realizing that this expository device is a gross simplification.

The Psychologists' Random Utility Model

Psychologists from at least the time of Fechner (1860) had observed imprecision and variability in human judgments of physical phenomena such as the weight of objects. Thurstone, in an attempt to model such behavior and also to move into the realm of preferences, proposed what we now know as the first random utility model. His model is essentially as follows:

$$U_{ip} = V_{ip} + \varepsilon_{ip} \quad (26)$$

where for item i and person p , U_{ip} is a momentary position along the person's judgment continuum,⁵⁰ V_{ip} is the expected value of the item along the continuum, and ε_{ip} is an independently and randomly distributed error with zero mean and variance Δ^2 .⁵¹ This psychological theory assumes that there exists a continuum on which V_{ip} is the person's "true" value for the object. When making a judgment response involving that object, however, the person uses a heuristic and pragmatic process that samples U_{ip} from a dispersion of ε_{ip} about V_{ip} , with ε_{ip} attributed to random fluctuations in perception as well as in judgment.

The psychologists' early model was not intended to interpret judgment as a function of the characteristics of the item (for example, size, color, price) or of the person making the judgment (for example, income, education, attitudes). Rather, it was developed to interpret variation in repeated responses involving the same item. The only way to estimate the "true value," V_{ip} , of item i to person p was to—in a clever way so as to obtain independent responses—repeatedly ask the judgment question and take a measure of central tendency from the set of individual responses.

Perception and judgment were assumed to be imprecise processes, but with the imprecision creating an orderly, stochastic dispersion. Thurstone (1927a) assumed a normal distribution for ε .⁵² He developed the law of comparative judgment and associated methods for interval scaling of paired comparison responses (Torgerson 1958). Dealing with preferences and lacking a dimension with an absolute zero point (such as money offers to economic studies), an interval-level scaling was sufficient and appropriate.

The RUM Model

Economists and others adopted this random utility notion, but for both theoretical and practical reasons thought of the utility that is implied or revealed by a consumer's choice as being precisely known by the consumer, not as a stochastic variable.⁵³ In other words, the economist assumes direct revelation of the psychologist's V and thus ignores the possibility of the psychologist's ε .⁵⁴

The economists' goal was different from the psychologists'. First, the economists aimed to explain utility as a function of measurable variables. Further, because of a preference for observed behavior (revealed preference), economists typically dealt with data containing judgments or behavior of many people, each of which usually provided only one observation, and wished to explain variation across people. To the economist then, in an application of the RUM model, the utility of an item to a person is as follows:

$$Z_{ip} = \mathbf{W}_{ip} + \eta_{ip} \quad (27)$$

where Z_{ip} is a precise/accurate indication of preference/utility of item i to person p , $\mathbf{W}_{ip}$ represents vectors of variables describing the item and the person, and ε_{ip} accounts for variation that is not captured—but is potentially capturable—in $\mathbf{W}_{ip}$. Thus, the original sense of ε (which included random variation in an individual person's choices due to errors of perception and judgment) is lost in the RUM model and replaced with variation due to unincluded variables in explaining variation across people or goods. The assumed precision on the part of the consumer was considered necessary to the economist's model of utility maximization, and, in any case, it greatly simplified computation, for it allowed for use of the logit model (McFadden, 2001).

Clearly the nature of $\mathbf{W}$ in equation 27 is vastly different from that of V in equation 26. $\mathbf{W}$ represents a potentially large set of characteristics, whereas V is a single expected value. $\mathbf{W}$ is a set of explanatory variables, whereas V is the key variable to be estimated. To make this distinction clearer we may rewrite equation 27 with a more explicit characterization of $\mathbf{W}_{ip}$, as follows:

$$Z_{ip} = f_{ip}(\mathbf{x}_i, \mathbf{y}_p) + \eta_{ip} \quad (28)$$

where $\mathbf{x}$ is a vector of variables describing the item and $\mathbf{y}$ is a vector of variables describing the person.

A Reconciliation of the Two Random Utility Models

Common sense and a preponderance of experimental evidence dating back many decades support the existence of human errors in perception and judgment. The assumption of the source of error in the RUM model is, we would argue, too limiting and a barrier to development of a realistic model of human behavior. A reconciliation of the two models is in order.

To review, the economist assumes that maximization of the instantaneous utility of an object (Z_{ip}) as implied by observed choices will maximize the chooser's welfare. The psychologist assumes that maximization of the person's true utility for the

object (V_{ip}) will maximize the chooser's welfare. In the psychologist's model, the chooser does not know his/her true utility for the object and imperfectly estimates V_{ip} in drawing an instantaneous U_{ip} from a random distribution about V_{ip} when making a choice. If we allow that utility is really stochastic, we must replace Z of equation 28 with $U = V + \varepsilon$ of equation 26, yielding:

$$V_{ip} = f_{ip}(\mathbf{x}_i \mathbf{y}_p) + \eta_{ip} - \varepsilon_{ip} \quad (29)$$

That is, η accounts for unincluded but potentially includable variables and ε accounts for uncludable (internal perceptual and judgmental) variables.⁵⁵ The linear combination of two random variables is another random variable such that $\lambda_{ip} = \eta_{ip} - \varepsilon_{ip}$, so that equation 29 becomes:

$$V_{ip} = f_{ip}(\mathbf{x}_i \mathbf{y}_p) + \lambda_{ip} \quad (30)$$

where λ is a composite error that is the difference between two randomly distributed processes.

If, in an actual application, the observed distribution of the random error (λ) reasonably approximates the distribution of $\eta - \varepsilon$, the RUM model actually aims at maximization of V_i as in equation 30. The chances of this are not unreasonable. Indeed, data in the form of judgments from numerous people with similar outward characteristics but different instantaneous utility (and associated judgments or choices) will mirror the stochastic behavior of individual choices. That is, homogeneous groups of people within the sample may approximate repeated judgments from one person. With enough data from each homogeneous group of the population, a RUM model may quite accurately estimate V . Acknowledging the existence of ε thus puts additional emphasis on sampling and careful, thorough data collection.

In summary, as we see it, all economic data sets contain random error reflective of ε as well as of η but, with a sufficiently rich data set, variability in individual human perception and judgment need not preclude estimation of V . Empirically, it seems that recognition of ε has little impact on a RUM study, given that rich data sets have always been the ideal. In the world of theory, however, work remains to fully incorporate ε into the economists' world.

Frequently Used Abbreviations and Symbols

Abbreviations

DS = double-sort (for double-sort procedure)
IIA = independence from irrelevant alternatives
IID = identically and independently distributed
MPS = mean preference score
PC = paired comparison (for PC program)
PCS = paired comparison simulation (for PCS program)
PSD = within pair preference score difference
RUM = random utility maximization

Symbols

f_{rc} = a cell in the frequency matrix giving the number of times the item in column c was chosen above of the item in row r
 a_c = preference score of the item in column c for a given respondent
 c = a column of a matrix
 C = number of circular triads
 i = an item to be compared
 j = an item to be compared
 M = Maxwell scale value
 p = a respondent
 r = a row of a matrix
 s = number of respondents
 t = number of items in the choice set
 T = Thurstone scale value
 U = instantaneous value of an item along the judgment or preference dimension of interest
 V = expected value of an item along the judgment or preference dimension of interest
 Δ = discriminial dispersion standard deviation
 $\mathcal{E}$ = error representing discriminial dispersion about V
 ζ = coefficient of internal reliability
 ρ = a proportion
 π = probability

Endnotes

- 1 Bock and Jones (1968) and Davidson and Farquhar (1976) cite other papers in German from the 1800's.
- 2 For this reason the method has been called a multiple-good valuation method (Brown and Peterson 2003).
- 3 We will assume throughout that the items being compared are equally commensurable on the specified choice dimension, and thus classes of items do not differ in the ease with which they can be compared, a possibility that was investigated by Tversky and Russo (1969).
- 4 Not offering an indifference option would be more of a problem if individual respondents were asked only one paired choice question, as opposed to a whole series of choice questions as is the case with paired comparisons.
- 5 We interpret Luce and Suppes as indicating that their model is particular to the individual person—that a given individual has a constant value for a given item.
- 6 The full equation for the variance of the difference between two normal distributions is $\Delta_{(u_i+u_j)}^2 = \Delta_i^2 + \Delta_j^2 - 2\gamma_{ij} \Delta_i \Delta_j$ where γ_{ij} is the correlation coefficient. Assuming independence, $\gamma_{ij} = 0$, and the term on the right drops out. Also, assuming identically distributed error terms, the variance reduces to $2\Delta^2$.
- 7 Intransitivity from the first cause has been called stochastic intransitivity, and intransitivity from the second and third causes has been called systematic intransitivity.
- 8 Many studies have demonstrated intransitivity in choices among multi-attribute items (see Rieskamp and others 2006, for a review). A prominent hypothesis for why such intransitivities occur is that people tend to pay more attention to those attributes for which the values differ substantially, de-emphasizing attributes that do not as dramatically distinguish the options. When different choices lead people to focus on different attributes, intransitivity may result.
- 9 Luce (1959) proposed that the IIA property can be viewed as a probabilistic version of the concept of transitivity.
- 10 Recent advancements in random utility modeling, such as the mixed logit model, address this shortcoming (for example, see Chapters 15 and 16 of Hensher and others 2005).

- 11 Kendall and Smith (1940) called their measure the “coefficient of consistence.” Because we are using the term “consistent” (or “inconsistent”) to refer to an unobservable measure, we have renamed the coefficient.
- 12 The difference between the two formulas arises because of difference in the numerical properties of odd- and even-numbered binary choice matrices. In an odd-numbered matrix it is possible for all the column (or row) sums to be equal such that the sum of the squares of the deviations of the sums from their mean is zero. That is the condition that maximizes the number of circular triads. It is not possible for all of the sums to be equal in an even-numbered matrix and it is therefore not possible for the sum of the squares of their deviations from their mean to equal zero.
- 13 However, the probability of a consistent choice does not necessarily remain constant over the course of a paired comparison session. Rather, as will be seen in Chapter 5, the probability of a consistent choice tends to increase over the course of the presentation of the pairs. This dependence of later choices on earlier ones—which is interpreted as a narrowing of $\mathcal{E}$, suggesting preference refinement—indicates that strict independence among choices does not hold.
- 14 Thurstone (1927a) presented the progressively more stringent assumptions as Cases I to V. Case I introduces some constraints on the correlations among the dispersions and Case III goes further and sets them to zero, establishing independence among the error terms. Case IV limits the differences among the dispersions and Case V goes further and makes them identical. Others later showed that the assumption of zero correlations for Case V could be relaxed to one of equal correlations (see Torgerson 1958).
- 15 As Ben-Akiva and Lerman (1985) explain, the way Thurstone characterized the problem was in fact a probit model, but Thurstone lacked the computing capacity to estimate the model.
- 16 Thurstone’s Case II suggests that paired comparison judgments from a whole set of respondents who each judge each pair only once may be similar to asking just one respondent to repeatedly judge the pairs of items. This case thus introduces the possibility of a homogeneous group of respondents.
- 17 The assumption that all respondents in a group have identical V s for an item is not so difficult to accept if the V s represent judgments of physical phenomena such as weights of objects, but it is more difficult to accept when the V s represent preferences. With preferences from a group of respondents, tests for differences among distinct subgroups are an important step.

- 18 The probability of an inconsistent choice when comparing items with V 's of 10 and 11 is 0.362 given $\Delta = 2$ (top graph of fig. 1-5) and 0.236 given $\Delta = 1$ (bottom graph). The corresponding probabilities comparing items with V 's of 10 and 16 are 0.017 and <0.001 .
- 19 Increasing certainty about one's preferences may be associated with learning or value discovery (Plott 1996) or even a value construction process (see generally Gregory and others 1993), whereby the act of considering various pairs of items causes the respondent to gradually reach a better understanding of where the items fall along the given dimension.
- 20 The distribution underlying the logit model, however, does roughly approximate the normal distribution. Thus, the logit is similar to the probit, but with the advantage that computation is easier.
- 21 Concerns about the degree to which a true interval scale measure is achieved become moot if anchors of generally agree-upon value on the dimension of interest are included among the items to be judged. For example, if monetary amounts are included among the items and it is assumed that monetary amounts have similar meaning to all respondents (technically, if income effects are absent), and if the choice dimension is, say, preference, ratio scale values of the goods can be approximated from the individual vectors of preference scores (Peterson and Brown 1998, Rosenberger and others 2002).
- 22 More recent studies that have used the PC program, not discussed herein, include Brown and others (2002), Nannini (2003), Rosenberger and others (2003), Brown and others (2005), and Brown and others (2008).
- 23 The image size for jpg photos is $W = 0.973$ inches (292 pixels), $H = 0.847$ inches (254 pixels), resolution = 300 pixels/inch. Photos (jpgs) for new monetary amounts may be created in Photoshop using DOLLAR AMOUNT TEMPLATE.PSD as follows: (1) open the psd file, (2) select Text tool (click on the top layer, or just type T), (3) double-click on text, (4) enter the new amount, click OK, (5) choose File.....Save a copy.....enter new name, and (6) in Save As, select jpg as the file type, maximum.
- 24 The key information from these matrices is presented in the Pairs.csv output file. The choice matrix output files were originally generated to facilitate debugging the PC program. They have been maintained in the final version of the program (PAIRCOMP6.EXE) to provide the user with visual evidence of some of the steps taken to produce the Pairs.csv file.
- 25 Another measure of inconsistency, one that provides a measure for individual items, is the number of circular triads in which an item appears (Hendel 1977). This reveals the frequency with which an item is involved in circular triads, but does not isolate individual choices that are most likely cause the circular triads.

- 26 It ought to be possible to explore these issues directly using the normal distribution theory on which the law of comparative judgment is based. Indeed, our investigation of the effectiveness of the DS procedure in identifying internally inconsistent choices began with an effort to develop a closed-form mathematical theory. That effort improved our understanding of the choice process, but the mathematics proved to be overwhelmingly complex because of the many stochastic variables varying simultaneously. See Paul Patterson (2003) “Research Done on Choice Matrices,” 03/22/02, for details on our effort to assess the DS procedure mathematically.
- 27 Providing subjects with an indifference option may have its advantages. If it worked well, it would allow us to separate real indifference from other sources of inconsistency. However, providing an indifference option runs the risk of allowing subjects to avoid all close calls. If, as Torgerson (1958) argues, the probability of true indifference at any given time is “vanishingly small,” forcing a choice maximizes the amount learned while still allowing indifference to be revealed in data from multiple subjects or from a given subject responding multiple times.
- 28 An ascending order for the dollar magnitudes in PAIRCOMP.TXT was necessary for the PC program to correctly model the assumed choices between dollar amounts.
- 29 MPS_c is also calculated as $\sum_{p=1}^s a_{cp} / s$ where a_{cp} for a given respondent p is as defined in equation 1.
- 30 Overlap becomes less likely if some of the items are far from the others on the dimension of interest (in other words, if the items fall into clusters). Clustering is avoided if the items are uniformly distributed along the dimension. Of course, one cannot know ahead of time how the items fall along the dimension, but analysts should at least avoid obvious cases of uneven distribution.
- 31 If appropriately spaced dollar magnitudes are included in the set of items and the purpose is to estimate dollar values for target items in the set, the non-linearity will distort the estimated value if the preference scores are assumed to comprise an interval scale anchored by those dollar magnitudes.
- 32 It is essential to note that the conditions under which we reached these conclusions for evenly and randomly spaced V s—IID normal discriminial dispersion and a large sample of respondents—are highly unlikely to occur in any practical application. Those conditions are, however, not unlike the assumptions implicitly embedded but too often ignored in estimation of RUM models from discrete choice contingent valuation and conjoint analysis experiments.

- 33 The change from the earlier Δ s of 1, 5, and 10, used for earlier illustrations, to the Δ s of 1, 2, 6, and 11 used here was made to explore additional questions not at issue here, and is of no consequence for the present analysis.
- 34 Within a cluster, the range in MPS varies with Δ . At $\Delta = 0$ the range in MPS across the seven items in the cluster would be 6. As Δ increases, the range diminishes. At $\Delta = 1$, the range is 2.2 (4.1 to 1.9, table 3-5) and at $\Delta = 2$, the range is about 1.4 (fig. 3-4).
- 35 With no interaction between proximate clusters, the MPSs of the items of the middle cluster would center at MPS = 11. However, because of some minimal interaction between the middle and the upper clusters, the middle cluster MPSs center at about MPS = 11.5 (fig. 3-5). This minimal interaction apparently causes the slopes of the two upper clusters to align linearly (a nearly straight line can be drawn through the points of the two upper clusters, fig. 3.5).
- 36 From the psychologist's perspective of the law of comparative judgment, we would expect the dollar magnitude to have discriminial dispersions with small, but non-zero standard deviations. Economic consumer theory would argue, however, that the dollar magnitudes are the numeraire and ought to be taken on face value with no dispersion whatever.
- 37 This result would be expected given the result presented in figure 3-4b, as the average Δ in the heterogeneous- Δ case is 3 versus a Δ of 2 for the figure.
- 38 The T scale cannot be computed with only one respondent. The R^2 of V to MPSs or to M scores averages 0.91 with only one respondent.
- 39 We assume that the probability that two items in a set of items will be perfectly tied in true preference order is vanishingly small. In other words, ties among items in expected preference order cannot occur.
- 40 One way out of this bind would be to obtain binary choice responses from a sufficiently large homogeneous sample of respondents (in other words, respondents with identical V s and discriminial dispersions). From such a sample, parametric estimates of the V s could be derived. Degree of consistency could then be estimated by comparing an individual's binary choices with the estimated V s of the items in each pair. Unfortunately, perfect homogeneity is extremely unlikely. Of course, the binary choices of a heterogeneous sample of respondent could be used to estimate aggregate V s for the items. Consistency, however, is a question of an individual respondent's V s, not a heterogeneous sample's V s, so we must resort to other means.
- 41 In the simulations we emphasize a Δ of 2 because, as is reported in detail further on, results from real-people experiments suggest that respondents' choices reflect a Δ of about 2.

- 42 Given $\Delta = 10$ instead of 2, the 100 percent threshold increases to about a PSD of 11. Also, with $\Delta = 10$, and ignoring ties, the DS procedure correctly found 69 percent of the inconsistent choices. Including ties, the procedure correctly found 73, 67, and 62 percent of the inconsistent choices for the ascending, random, and descending input orders, respectively. The proportion correctly found is greater at $\Delta = 10$ than at $\Delta = 2$ because higher Δ s produce more inconsistencies of higher PSD and the DS procedure is more effective at higher PSD.
- 43 We also examined the performance of the DS procedure when the V s are randomly spaced within three randomly spaced clusters such that the spacing within clusters is much smaller than the spacing between clusters. We found that clustering impairs the performance of the procedure and should be avoided.
- 44 Randomly assigning Δ s ranging from 1 to 5 results in an average Δ of about 3. This is a greater average dispersion standard deviation that is implied by comparing simulations to results of real-people experiments, which is close to 2 (see Chapter 4).
- 45 For these runs, the range in V is restricted in each case to the number of items (for example, for the run with 10 items, the items ranged in V from 1 to 10). Thus, given a constant Δ of 2, the proportion of choices that are inconsistent decreases as the number of items increases because the number of choices that are inconsistent increases much more slowly than does the total number of choices as the number of items increases (this is because the overlap among discriminial dispersions does not change much with changes in the number of items).
- 46 Similarly, we found that Δ in real-people experiments with 21 items is approximately 2.0. For the simulation with 21 items and $\Delta = 2$, mean $\zeta = 0.94$ and the mean proportion of choices that are consistent = 0.92.
- 47 Figure 5-10 shows a time of about 32 seconds for the first choice. This is an over-estimate caused by an error in the PC program that caused the time for the first choice to include the time used to view the previous screen, which contained instructions on how to respond to the choices to follow. The instruction screen may have taken the typical respondent roughly 5 to 15 seconds to view. This problem has been corrected in PAIRCOMP6.TXT.
- 48 The greater scatter for inconsistent choices than for consistent choices (shown in fig. 5-11) is probably due to the relatively lower number of cases among the inconsistent choices.

- 49 One possible criterion borrows an elasticity concept from economics. The axes in figure 5-14 both vary from 0 to 1 as a proportion. One might set the threshold at the point on the curve where the marginal decrease in ζ attributed to the next less consistent respondent becomes greater than the marginal increase in the proportion of the sample caused by that respondent. Such a rule is as arbitrary as any, but at least it could be applied consistently.
- 50 We will use “judgment” to include preference.
- 51 Earlier psychophysical models may have thought of V as the absolute magnitude of the item on the dimension of interest (for example, weight, brightness), but when Thurstone moved into the domain of preference, he allowed for individual variation in V because V incorporated matters of taste and preference that were assumed to vary across individuals.
- 52 Thurstone called the outcome a probit model but he lacked the ability to estimate it (Ben-Akiva and Lerman 1985, p. 128).
- 53 Hanemann and Kanninen (1999, p. 307) make precisely this point: “In a RUM model it is assumed that...the individual knows her preferences with certainty and does not consider them stochastic...”
- 54 Ben-Akiva and Lerman (1985) identify two probability choice models which they call “constant utility” and “random utility.” The constant utility model is consistent with a psychological theory of choice, with the following caveats: the “constant utility” of an object is “fixed” for a given person but may vary across persons and the “fixed” utility within the person is the expected value of the instantaneous or apparent utility implied by a specific choice. The random utility model is consistent with the microeconomic utility maximization theory of choice. Thurstone (1927a) and others applied the utility theory underlying the psychological choice model to measure perceptions, judgments, and preferences on an interval scale from responses obtained by paired comparison and categorical rating scales. The economic RUM model has been developed and applied by economists and others to predict choices and estimate and explain implied economic values.
- 55 If one thinks of perception and judgment as fully explainable in terms of measurable variables (for example, measures of individuals’ brain functions), such that the meaning of y_p is expanded to include as yet unmeasurable variables, the psychologists’ model becomes unnecessary and the RUM model suffices. Recent advances in neuroscience notwithstanding, we are probably a long way from the day when individual variation in choice is deterministically modeled.



The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of the National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals. Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Station Headquarters

Rocky Mountain Research Station
240 W. Prospect Road
Fort Collins, CO 80526
(970) 498-1100

Research Locations

Flagstaff, Arizona	Reno, Nevada
Fort Collins, Colorado	Albuquerque, New Mexico
Boise, Idaho	Rapid City, South Dakota
Moscow, Idaho	Logan, Utah
Bozeman, Montana	Ogden, Utah
Missoula, Montana	Provo, Utah

The U.S. Department of Agriculture (USDA) prohibits discrimination in all its programs and activities on the basis of race, color, national origin, age, disability, and where applicable, sex, marital status, familial status, parental status, religion, sexual orientation, genetic information, political beliefs, reprisal, or because all or part of an individual's income is derived from any public assistance program. (Not all prohibited bases apply to all programs.) Persons with disabilities who require alternative means for communication of program information (Braille, large print, audiotope, etc.) should contact USDA's TARGET Center at (202) 720-2600 (voice and TDD). To file a complaint of discrimination, write to USDA, Director, Office of Civil Rights, 1400 Independence Avenue, S.W., Washington, DC 20250-9410, or call (800) 795-3272 (voice) or (202) 720-6382 (TDD). USDA is an equal opportunity provider and employer.