1996

Value Measurement for New Product Category: a Conjoint Approach to Eliciting Value Structure

Roland Helmut Heger
Portland State University

10.15760/etd.1304

Value Measurement For New Product Category:
A Conjoint Approach To Eliciting Value Structure

by

ROLAND HELMUT HEGER

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY
in
SYSTEMS SCIENCE: BUSINESS ADMINISTRATION

Portland State University

DISSERTATION APPROVAL

The abstract and dissertation of Roland Helmut Heger for the Doctor of Philosophy in Systems Science: Business Administration were presented April 17, 1996, and accepted by the dissertation committee and the doctoral program.

COMMITTEE APPROVALS: [redacted]
Robert R. Harmon, Chair

[redacted]
Thomas R. Gillpatrick

[redacted]
George G. Lendaris

[redacted]
Richard W. Sapp

[redacted]
Dundar Kocaoglu
Representative of the Office of Graduate Studies

DOCTORAL PROGRAM APPROVAL: [redacted]
Beatrice T. Oshika, Director
Systems Science Ph.D. Program

* *

ACCEPTED FOR PORTLAND STATE UNIVERSITY BY THE LIBRARY

[redacted] on 22 May 1996

ABSTRACT

An abstract of the dissertation of Roland Helmut Heger for the Doctor of Philosophy in Systems Science: Business Administration presented April 17, 1996.

Title: Value Measurement For New Product Category: A Conjoint Approach To Eliciting Value Structure

Ability to measure value from the customer's point of view is central to the determination of market offerings: Customers will only buy the equivalent of perceived value, and companies can only offer benefits that cost less to provide than customers are willing to pay. Conjoint analysis is the most popular individual-level value measurement method to determine relative impact of product or service attributes on preferences and other dependent variables.

This research focuses on how value measurement can be made more accurate and more reliable by measuring the relative influence of selected methodological variations on performance in prediction and on stability of value structure, and by grouping customers with similar value structure into segments which respond to product stimuli in a similar manner. Influences of the type of attributes included in the conjoint task, of the factorial design used to construct the product profiles, of the type and form of model, of the time of measurement, and of the type of cluster-based segmentation method, are evaluated.

Data was gathered with a questionnaire that controlled for methodological variations, and with a notebook computer as the measurement object. One repeated measurement was taken.

The study was conducted in two phases. In Phase I, influences of methodological variations on accuracy in prediction and on respective value structure were examined. In Phase II, different cluster-based segmentation methods — hierarchical clustering (HIC), non-hierarchical clustering (NHC), and fuzzy c-means clustering (FUC) — and according conjoint models were evaluated for their performance in prediction and in comparison with individual-level conjoint models.

Results show the best models for a variety of design parameters are traditional individual-level, main-effects-only conjoint models. Neither modeling of interactions, nor segment-level conjoint models were able to improve on prediction. Best segment-level conjoint models were obtained with a fuzzy clustering method, worst models were obtained with k-means and the most fuzzy clustering approach.

In conclusion, conjoint analysis reveals itself as a reliable method to measure individual customer value. It seems more rewarding for improvement of accuracy in prediction to apply repeated measures, or gather additional data about the respondent, than to attempt improvement on methodological variations with a single measurement.

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

List of Tables

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER I

INTRODUCTION

From a marketing perspective, the ability to measure value from the customer´s point of view is fundamental to the development of successful marketing strategy. Knowledge of customers´ value structure allows one to compare product/service benefit components created by diverse company activities, and allows for pricing that captures some of the value content of a product in form of sales proceeds. Thus, *value measurement is central to the determination of market offerings*: Customers will only buy the equivalent of perceived value, and companies, in the long run, can only offer benefits that cost less to provide than customers are willing to pay.

While it is relatively easy to gather data about *aggregate* buying behavior, the structure of the decision process exhibited by the *individual* buyer is difficult to reveal. The same is true with respect to familiar and unfamiliar product categories. Several competing methodologies, favored by researchers with differing perspectives, have been in use to shed light on the components of customer value, reveal their interactions, and predict preferences and choice behavior by modeling the structure of the customer value system.

Researchers from many scientific disciplines have been tackling value measurement and choice: psychologists who have mainly been interested in the mental constructs determining evaluations and choice behavior, microeconomists whose interests have been focusing on the efficiency of market choice behavior, engineers who attempt to arrive at optimal designs with the use of value analysis, operations researchers (OR) and management scientists (MS) who have been modeling customer value and choice within normative frameworks of rational decision making, and marketers whose focus has been on models of value judgment that may be readily translated into actionable elements of the marketing mix. All of these disciplines favor different valid approaches and mathematical frameworks to quantify effects of value judgment.

Conjoint analysis, introduced to marketing by Green and Rao (1971), is a family of methods that enables marketers to determine the relative impact of product or service attributes on preferences and other dependent variables. It is enjoying increasing popularity, particularly in recent years (Wittink, Vriens, and Burhenne 1994, p. 41), because it provides a flexible framework for modeling and understanding value judgments at different levels of marketing decision making, at the individual customer´s level and at different aggregate levels. Additionally, it is useful for decisions about various elements of the marketing mix, linking customer perceptions and business objectives, as for instance utility measurement, buyer choice simulations, product design, pricing, market segmentation, and competitive strategy (Green and Krieger 1993, p. 468). In providing a means for modeling individual decision structures it offers the promise of greater insight into these structures with supposedly greater accuracy of measurement and potentially greater power for managers to influence the customer decision process.

1.1 Study Purpose and Problem Overview

A major problem in estimating customer value is that realistic value decision contexts in conjoint measurement tasks typically involve a relatively large number of product attributes with associated large numbers of responses required for estimation of customer value structure. With only limited numbers of responses per subject, and a large number of parameters to be estimated, reliability of individual-level models becomes doubtful. As many a conjoint study´s purpose is to elicit value structure for target segments of the market (Wittink, Vriens, and Burhenne 1994; Wittink and Cattin 1989; Cattin and Wittink 1982), researchers suggested and developed methods to aggregate respondents and estimate value structure on a segment-level basis. Segments derived with this approach are based on differences in individual benefit components, as opposed to common demographics as bases for segments. The suggestion in the literature to improve on prediction in conjoint with segment-level benefit derivation is pursued in this study, further. It is currently an area of intensive research.

The primary purpose of this research study is to measure the relative influence of

- selected methodological variations of conjoint analysis and
- segmentation methods (i.e. grouping of subjects)

on customer value structure. Influences of the type of attributes included in the conjoint task, and of the factorial design used to construct the product profiles are evaluated, on their own, and in their combination. In particular, some recently suggested and some new cluster-based segmentation methods in the context of conjoint measurement, non-hierarchical and fuzzy clustering respectively, are compared with each other, with traditional benefit clustering and segmentation procedures, and with traditional, individual-level based value estimation techniques, in order to determine the best approach for value estimation in connection with segmentation. A moderately complex, relatively new product category, a laptop or notebook computer, is chosen as the measurement object.

Many commercial and research-based conjoint studies conducted to date have explored value components for familiar product categories, as for instance apartments and cars, and were limited in model flexibility and estimation procedures. Recent suggestions to overcome these limitations with benefit segmentation approaches (i.e. subject grouping and estimation of benefit attributions on the group level) are applied and conjectured improvements in reliability and predictive validity are tested. Specifically, allowing for overlapping cluster segments by applying a fuzzy clustering algorithm enabled tests for potentially higher predictive accuracy compared with individual-level predictions and traditional grouping techniques. This provides some additional insights into the adequacy of specific conjoint model types and methodological procedures for differing marketing purposes, namely for prediction and segmentation. With key marketing decisions based on conjoint studies, empirical evidence of reliability and validity in differing contexts is of primary concern (Bateson, Reibstein, and Boulding 1987, p. 451).

This study, more specifically, addresses the following research questions:

1) What is the influence of the type of attribute chosen for the evaluative task (i.e. technical or product-referent attributes and non-technical or user-referent attributes) on customer value structure and predictive validity ?

2) What is the influence of specific factorial designs, i.e. of specific combinations of product attribute values, on estimation of customer value structure and predictive accuracy ?

3) How do type of attribute in the product profile and factorial design interact in their influence on customer value structure for different models ?

4) Which individual-level model for customer value structure performs best with respect to prediction ?

5) Can cluster-based segmentation approaches improve accuracy in prediction of value attributions to product profiles over individual-level conjoint models ?

6) Which aggregate model for customer value structure performs best with respect to prediction ?

7) Are the purposes of prediction and segmentation, as well as potential other purposes, better served with the suggested methods, and what practical limitations are there for the different methods to support specific purposes ?

8) Are benefit segments obtained with different clustering procedures meaningful for target marketing, or may they only increase predictive accuracy ?

1.2 Importance of Proposed Research

Value measurement has enjoyed considerable attention in recent years due to the following reasons:

- Accepting the marketing notion that the very rationale for companies to be in business is the satisfaction of needs and wants, the concept of value as well as the need to measure it is pervasive. Providing value to the (individual or industrial) customer is the basis of the marketing concept. It may affect all parts of a marketing program, and may form the foundation of corporate strategy (Green and Krieger 1993, p. 468).

- As further refined market segmentation, targeting, and positioning become increasingly important, so does the availability of corresponding value measurement instruments in order to be more accurate, effective, and efficient in performing these tasks. In the context of measuring consumer preferences for multiattribute products in product design, conjoint analysis enjoys widespread popularity. Accordingly, an effort to improve on its accuracy and applicability for different marketing and general business purposes, as for instance for corporate strategy or segmentation, is a potentially rewarding endeavour.

Value measurement with conjoint analysis is useful to the researcher and practitioner in several ways:

1. Knowing what customers value in an offering and to what extent, allows the marketer to *concentrate his resources* and perform activities that provide the best tradeoff between highest possible customer benefits and lowest possible costs. Specifically,

   a) in devising a product policy, the knowledge of how customers perceive value in product attributes and performance criteria may be used to combine those sets of product features that are most attractive, respectively have the highest perceived value, for specific customer segments. If segmentation is not done on the basis of a priori defined variables, value perceptions may very well be used to segment the market (cp. Laitamaki and Renaghan 1988, p. 179). Buyers of consumer electronics, for instance, may be classified into price-sensitive and into feature-oriented customers, depending on the importance of different dimensions of value (i.e. product attributes) to them.

b) In devising its promotional strategy, the vendor needs to understand where the value of the product is created, if value is added by emphasizing particular product features, or by the way the product is presented, by the packaging, the prestige that is associated with the product, or by the channel in which it is sold. Understanding the value components of an offer through appropriate value measurement enables the vendor to choose the most effective promotional policies for targeting and positioning.

c) In planning one´s distribution strategy, understanding customer values allows choosing distribution channels consistent with product and promotional policy for more efficient value delivery.

2. Knowing what value customers perceive in an offer allows the firm to *price its offer* according to these perceptions and the firm´s own objectives, be they to increase market share, to increase profits, or just to provide superior benefits at the lowest possible price. It allows goal-oriented pricing, goal-oriented capturing of perceived customer value, instead of arbitrary cost markups.

The central problem of marketers who want to address new markets seems to be a lack of understanding *what* exactly it is that customers value in an offer, and *how* customers form value perceptions about products and product categories. Therefore, the problem may be separated into at least two parts:

(1) The concept of customer value is not clearly and unambiguously specified. Neither is its relationship with product features or attributes. Attributions of benefits to product characteristics is at the heart of conjoint measurements. Knowledge of effects of the type of attribute on conjoint measurements may provide some clues as to the extent with which type of attribute influences value measurement, allowing for potential improvements in the measurement instrument. Specifically, the assumption of conjoint analysis of independence from irrelevant

attributes (IIA) may be violated by including abstract or user-referent attributes into the product profile (Oppewal, Louviere, and Timmermans 1994, p. 104).

(2) The theoretical framework for representation of value concepts and choice behavior, i.e. as a "deterministic" decision problem, as a "fuzzy" decision problem, or as a random choice problem, is also not obvious or unambiguous, though at first glance one might assume this problem context to favor representation as a rational, deterministic decision problem over the others.

Without improved methods for addressing these problems, there will likely be more product failures similar to one experienced by AT&T: On Thursday, July 28, 1994, The Wall Street Journal reported that AT&T will close EO, Inc., the subsidiary that developed and manufactured new electronic devices dubbed 'personal data communicators' or 'personal digital assistants' (PDAs), as sales of its EO Personal Communicator lagged considerably behind expectations (Naik 1994, p. B5). Industry analysts have identified palm-size devices that keep people organized and connected as a new, potentially huge market, "if only someone can come up with a device with the right combination of features — something no one has come close to doing yet" (Lee 1994, p. 6). Accordingly, the president of the AT&T Consumer Products unit, which will continue EO, Inc.'s development efforts, and an EO board member was quoted lamenting: "I wish we knew what customers wanted" (Naik 1994, p. B5). In order for "the right combination of features" to be decisive, however, user-referent (i.e. non-technical or non-physical) attributes must not influence the relative evaluations of physical attributes (i.e. features).

Value measurement in itself is a multivariate problem that entails a great deal of complexity, due to the fact that several variables have to be considered simultaneously rather than individually or sequentially. This study was intended to help better understand methods to identify measurable and actionable market segments that are likely to respond differently to various policy actions. It yields guidelines for operationalization of increased customer-orientation. In particular, one needs to know the relationship between purpose and a particular method, as for instance conjoint methods that rely on groupwise parameter estimates do not allow for individual differences in benefit structure any more, but they provide the advantage of significance testing of benefit components.

This study is the first empirical comparison of recently proposed aggregation methods to improve on

- reliability,
- predictive accuracy, and
- segmentation.

The only two limited studies (Green, Krieger, and Schaffer 1993a; Green and Helsen 1989) involving two newly suggested segmentation approaches in conjoint analysis could not confirm the respective authors´ claims of superiority of newly proposed methods over competing ones, specifically in predictive accuracy. It is also the first empirical study that intended to integrate examination of selected methodological manipulations of conjoint analysis *jointly* with segmentation approaches in a controlled manner in order to determine their relative separate and joint impact on dependent variables and surrogate performance measures. However, after completion of the first phase of research, examination of joint effects turned out to be of no relevance any more. Traditional grouping methods in connection with conjoint analyses are extended with

- non-hierarchical clustering methods (k-means), and
- fuzzy clustering (fuzzy c-means).

There is a marked absence of studies concerning effects of the nature of product category on estimation of value structure. Specifically, type (but also number) of attributes as well as their contextual settings (i.e. correlations and structural relations as for instance hierarchy among them, relevancy as attributes, familiarity of respondent with category) are seldom dealt with and often glossed over. However, exploring the effects of attribute variations is possibly more fundamental to the nature of value structure than methodological particularities, as for instance the estimation technique used, and the former is relevant for all methodological variants of conjoint analysis (see Elrod, Louviere, and Davey 1992, p. 376; Wittink, Vriens, and Burhenne 1994, p. 49: "...it would be useful to return to more fundamental questions that apply to all methods, such as the definition and choice of attributes and (number of) levels"; Oppewal, Louviere, and Timmermans 1994, p. 104). Obviously, more research on this subject is necessary. One of the main purposes of this study was to explore this effect.

Immediate applicability and practicality of suggested research to the conduct of conjoint analysis studies and subsequent elicitation of value structure for improved marketing decision making is obvious, but so is the danger of overconfidence in empirically untested methodological procedures and their respective results in decisions such as: what kinds of attributes to emphasize in product design and communication, segmentation according to "value" or benefit segments instead of a priori segmentation on the basis of demographics, pricing for different market segments, and similar questions. The issue and justification for researching methodological variations is greater ability in making informed choices for necessary tradeoffs among problems addressed and capability of respective methods. Such investigation allows for better fine-tuning of methodological choices to the problems at hand.

Finally, with conjoint studies forming the basis of substantial, long-term business decisions (Green and Krieger 1993, p. 468), knowledge about the method's capabilities and limitations is of utmost importance in order to carefully balance strategic decisions and ensure predictability of new product success.

With its combination of experimental manipulations (see section 3.3.1 of Chapter III), this study provides an important contribution to advance theory of consumer decision making, particularly in value measurement, concerning product and attribute type, decision model (additive or interactive), estimation and validation approaches, and the usefulness of value measurement techniques for two specific marketing purposes: prediction and segmentation.

Both parts of the general problem in value measurement, i.e. what customers value, and how they form value perceptions, have repercussions on value measurement, i.e. what is measured and how measurements are taken and subsequently processed. Additionally, these measurement problems occur within broader frameworks of customer choice behavior and respective modeling attempts which may be attributed to the background of the researcher and its perceptual and analytical stance, to the nature of the problem, and its context or purpose.

1.3 Research Perspective and Decision Model

The perspective taken for this study is heavily influenced by the systems view: Problem, context, and perceiver are mutually dependent, and it is the perceptual stances taken that admit for the notion of system and allow appropriate focus on subject matter (see Lendaris 1986). Value judgments occur at different system levels, with each system being defined on the former three dimensions (i.e. problem, context, and perceiver). One value system, for instance, may be defined by one product with its respective attribute combination, customers who evaluate the product according to some behavioral characteristic towards it, for instance a preference judgment, and the researcher who is interested only in the preference judgment to measure the product´s utility for these customers for prediction. The researcher acts as systems analyst of the customer´s value system. The researcher may assume other, different roles as an analyst on the same system level, or on the suprasystem level, for instance as a business strategist using customer value structure as primary input (for the relevance of systems levels and the perceptual stance of the perceiver see Lendaris 1986). In a different context, the researcher may pursue several objectives in observing the system, for instance product policy and market share determination, but he may exercise the same pattern of assuming analyst´s roles and perceptual stances (i.e. systems levels). The mathematical models used to represent customers´ judgmental systems are as diverse as the perspectives taken, and the behaviors exhibited by individuals.

The perspective taken in this study is based on the assumption that individuals evaluate products and services by integrating salient attribute information about them. This entails the view of a process-oriented model of consumer decision making, in contrast to stochastic choice models (Lilien, Kotler, and Moorthy 1992, pp. 31 and 56). Stochastic choice models do not attempt to explain or predict choice by modeling its supposed determinants — mental constructs and/or product attributes. Instead, value judgments are viewed as results of a process that has no discernible pattern beyond that which can be explained by assigning purchase probabilities for the available alternatives. These models are particularly appropriate for low-involvement products, as for many beverages and food products, where little conscious decision making takes place. This study did not consider such situations.

Process-oriented models assume underlying customer and/or product related determinants of value judgments which can explain large portions of choice variations in a deterministic manner. These models were often found to be more appropriate for high-involvement decisions as with purchases of durables or investment type goods. A helpful characterization of judgmental phases is the well-established assumption of up to five stages in the consumer decision process which borrows from ideas in the OR/MS and problem-solving literature about rational decision making. This assumption is based on the theory of information integration (Anderson 1981,1982) and attitude research. The five stages are: need arousal, information search, evaluation, purchase, and postpurchase feelings. In the quest to design parsimonious but empirically valid decision models, different stages of the customer´s (decision) process offer varying degrees of insight in different consumer decision situations and for different managerial problems. The stages´ appeal for marketers lies in the fact that each stage suggests different possibilities to facilitate or influence the decision process, as well as measurements to be taken to calibrate a model and aid in decision support. This study focused on evaluative processes that lead to choice, and Figure 1 on page 14 illustrates its context.

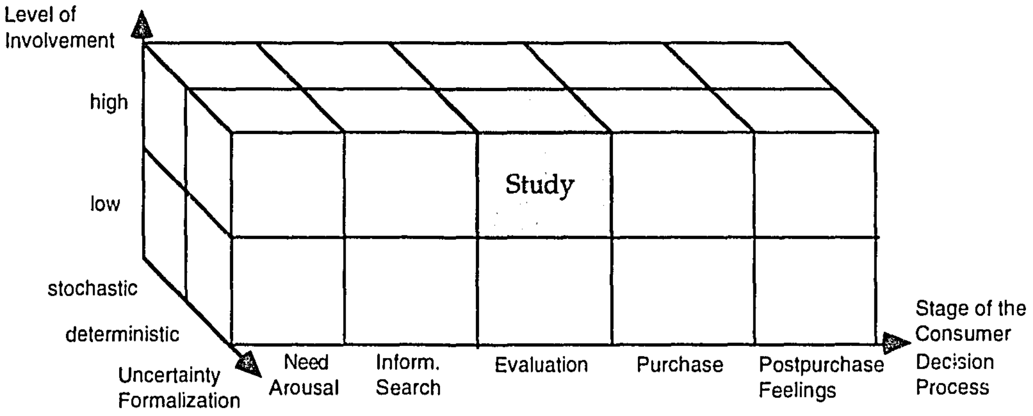


Figure 1. Decision Context of Study.

Developing the framework further, evaluations may be partitioned into two components: First, beliefs about the presence or absence of attributes and their magnitude in product alternatives (i.e. perceptions) must be established. Second, based on those perceptions, attitudes towards alternatives (i.e. preferences) must be determined which indicate how favorably disposed people feel towards the alternatives, most often with the ultimate goal to predict purchase likelihood. In the OR/MS literature, the first component is often assumed to be easily agreed upon among different perceivers and, thus, not explicitly modeled, while such differences in perceptions tend to be the focus of psychologists. Marketers vary between these two extremes in foci. Perceptions and preferences are often assumed to be respective phases of the evaluation process, but this need not necessarily be the case. Figure 2 on page 15 illustrates dimensions that help structure the evaluation task, and respective models should be able to incorporate these dimensions.

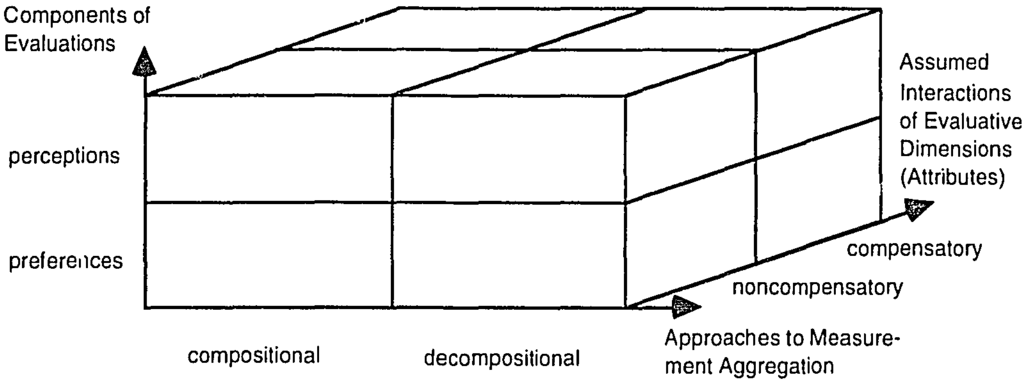
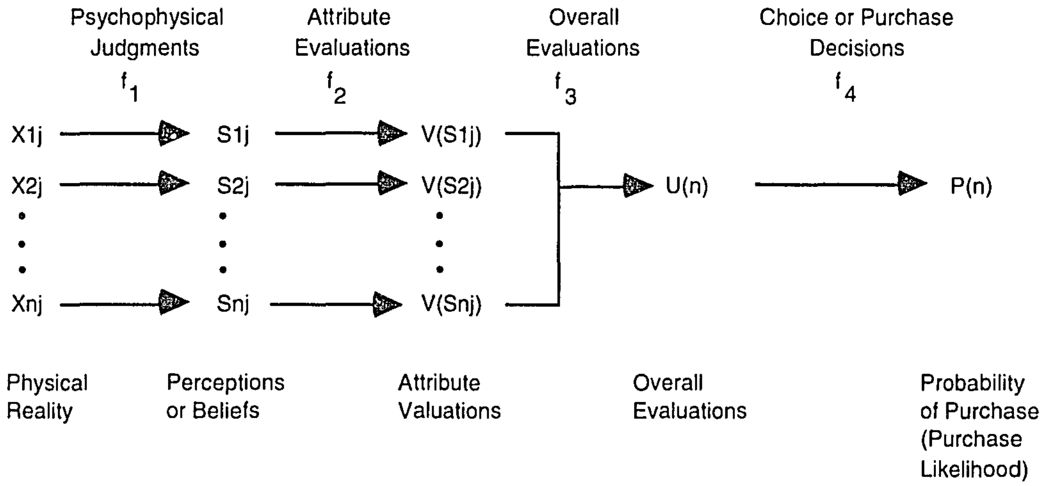

Components of Evaluations
perceptions
preferences
compositional
decompositional
Approaches to Measurement Aggregation
Assumed Interactions of Evaluative Dimensions (Attributes)
compensatory
noncompensatory

Figure 2. Dimensions of Evaluation Models.

Perceptions can be measured directly by asking customers how much of an attribute they perceive a certain product to contain (i.e. compositional and self-explicated approach), or they can be inferred by asking how similar certain products are and then deriving what discriminates between different products (i.e. decompositional approach). The general idea behind the derivation of evaluative criteria and perceptual dimensions is that customers commonly do not use product attributes directly for comparisons but distill their perceptions into a limited number of high-level, abstract evaluative dimensions. Usually, customers cannot express these dimensions directly. The most frequently used methods for the compositional approach to measuring perceptions are based on factor analysis (FA), and those for decompositional methods are based on multidimensional scaling (MDS).

Attitudes are measured analogous to perceptions by asking customers about their preferences concerning attributes (i.e. compositional and self-explicated approach) or by asking them about preferences among alternatives and deriving preferences among attributes (i.e. decompositional approach). Most models of attitude formation assume that choice behavior, as well as attitudes, are determined by judgments on specific attributes of the choice object, expressible in form of a utility or expectancy value function. Therefore, they transform judgments based on attribute evaluations (i.e. attribute levels and importance) to a single-dimensional scale of brand attitude or product utility, frequently after first distilling the attribute information into higher-order decision factors using perceptual mapping techniques. The applied mapping or aggregation techniques model assumed variations in customer valuations with alternative combinations and levels of attributes. The most frequently used methods are multiattribute utility (MAUT) functions in form of compensatory and noncompensatory models, with conjoint analysis (ConjA) being the most popular MAUT method. An alternative approach is structural modeling of preferences (Bagozzi 1982). In a compensatory model the weakness of a product alternative on one attribute can be compensated for by strengths on another, and the attributes are summed to determine the favorability or unfavorability of the attitude towards the product (e.g. the Fishbein model; Ajzen and Fishbein 1980). This is the most common way of product evaluation. In noncompensatory models usually only a few attributes are used to evaluate a product, and shortcomings on any one attribute cannot be compensated by more favorable levels on another. This behavior is often found in stages prior to information processing, e.g. in information search. Various mixed, sequential rules may be used. Figure 3 on page 17 (adapted from Louviere 1988, p. 10) depicts the general model of decision making used in this study. It is detailed in section three of Chapter II.



(adapted from Louviere 1988, p. 10)

Figure 3. General Model of Customer Evaluation (Decision Making).

Conduct of value measurement with a conjoint study that approximates consumer judgment and decision processes as illustrated in Figure 3 on page 17 entails the following phases:

1. Gain an understanding of the decision problem and its environment faced by target individuals. It involves answers to the following questions:
   - What are the elements of utility for the product, service, or idea considered ?
   - What are the key decision criteria involved in the evaluation process ?
2. Design a conjoint experiment to understand how target individuals integrate decision attributes, i.e. how they evaluate multiattribute alternatives.
   - This involves specifying the type and number of attributes, as well as the attribute positions (or levels) used for construction of product profiles.

- It also involves specifying the basic model form of information integration, i.e. additive and interactive attribute terms, and linear, quadratic, or separate part-worths.
- Furthermore, it entails the factorial design used to create stimuli for evaluation, i.e. factorial versus fractional factorial designs, the presentation method for the profiles, selection of a preference measure, and selection of an estimation technique.

3. Identify measurable and actionable market segments that are likely to respond differently to different policy actions.

This study focused on influences of selected methodological choices, namely type of attribute and type of fractional factorial design on the individual level, as well as on segment-level value structure, and performance of respective models for prediction.

1.4 Definitions and Terminology

The following definitions, concepts and terms are used throughout the remainder of this study and may be illustrated by Figure 3 on page 17. Formal algebraic developments are provided in the methodology section of Chapter III.

Physical variable. This term refers to observations or measurements of various physical properties of the product or service considered. These properties are antecedents of determinant attributes in the theory of information integration (Anderson 1981, 1982).

Attribute. This is the term used for the determinant decision criteria customers are assumed to use to evaluate products or services (denoted by X_{nj} in Figure 3, where n is the number of the product alternative considered, and j is a subscript for the attribute).

Position (= Level). "Beliefs" that customers have about the amount of each determinant attribute possessed by products or services, called the "positions" on attributes (denoted by S_{nj} in Figure 3, where n is the number of the product alternative considered, and j is a subscript for the attribute). They are also referred to as levels when only discrete positions are considered.

Part-Worth. Judgment that a customer makes regarding "how good", "how satisfactory", or "how whatever" particular positions of particular products might be on particular determinant attributes. They are also referred to as part-worth utilities for the positions (levels) of those attributes.

Overall utilities (overall evaluations). Judgments, impressions, or evaluations consumers form of products and services, taking all the determinant attribute information into account. It is assumed that this evaluation is performed wholistically (in a Gestalt sense) and not holistically, but one makes the simplifying assumption that this judgment may be decomposed into its parts with reasonable approximate accuracy (for the distinction between wholistic and holistic perceptions of a system see Lendaris 1986, p. 605).

Brand. This term denotes a particular product or service available or possible on the market that can be evaluated and possibly selected by a customer, i.e. a choice alternative.

Final choice set. This is the set of brands a consumer seriously considers prior to making a choice. In marketing this choice set is often called the "evoked set".

Choice. This term refers to the cognitive process by which a consumer, after evaluating all the brands and forming a final choice set, decides to select one of the brands or not to make a choice.

New product category. This term is used for products that are satisfying new needs or wants, or satisfy established needs and wants only possible in the specific combination of features of the product.

(Attribute) Interaction. This term refers to the effect that the presence of particular levels of "other" attributes influences a particular attribute evaluation as well as the overall utility. In the presence of particular "other" attributes, utilities of one attribute´s levels may be diminished or increased.

Replication. One replication means one performance of a conjoint experiment, i.e. two replications denotes two performances of the experiment (not three).

1.5 Study Overview

First, an examination of the trade literature, interviews with sales reps and the manager of a local computer store, as well as a pretest of attribute importance for laptop or notebook computers yielded the attributes and their respective levels to be evaluated by respondents in this study. Information from the interviews was reconciled with secondary information about important product attributes, primarily published surveys and trade journal information. As the pretest did not reveal new, broadly important information about additional attributes, the set of product attributes and levels was obtained as conveyed in Table X on page 99.

An experiment was designed so as to address the research questions. Respondents were sampled from a medium-size university in the Northwest of the United States. The type of attributes included in the design of stimuli to be evaluated was varied systematically in order to determine their effect on estimation of part-worth utilities and their respective importance. Stimuli were presented as full-profile sheets of paper, and subjects were asked to rate their likelihood of purchase for respective product profiles. The number of attributes per stimulus were limited to nine attributes, in order to limit variations due to fatigue and information overload on part of the respondents.

Two fractional factorial designs were developed to test for effects stemming from different designs. The designs were devised so that limited two-way interaction terms are possible. Effects of inclusion of interaction terms are tested by estimating part-worth utilities with the inclusion of interaction terms, as well as without them in a merely additive model. Individual value structure is estimated with ordinary least squares regression (OLS).

Part-worth utilities are derived in two phases. In Phase I, individual-level part-worth utility models are derived and effects of variations in the type of attribute set (two dimensions) and in the type of fractional factorial design (two dimensions) are estimated. Additionally, in Phase II, three clustering procedures are performed to group respondents according to their part-worth structure into benefit segments: hierarchical clustering, a commonly used (hard) non-hierarchical cluster algorithm, and a new fuzzy cluster algorithm (fuzzy c-means) developed by Bezdek (1981, Chuah and Bezdek 1987) which allows for overlapping clusters with conjectured improved predictive accuracy.

Finally, three types of reliability and one type of validity are tested with various measures. Specifically, reliability over time (by administering two replications), reliability over attribute set (by varying type of attributes), and reliability over stimulus set (by employing two different fractional factorial designs) are tested. (Convergent) Validity is tested by comparing predictive accuracy of the conjoint models derived with self-explicated part-worths.

Due to the interdependencies of respondent tasks, study design, and methodological choices employed, it is necessary to restrict experimental variations to those that are not confounding each other´s effects. As is highlighted in most recent surveys of conjoint studies´ literature (Bateson, Reibstein, and Boulding 1987; Green and Srinivasan 1990) many studies vary too many parameters, confounding effects, and thus leading to contradictory results which have to be resolved later by conducting comparative studies with more focused, limited experimental manipulations.

Table I on page 23 summarizes the methodological variations of conjoint methods applied in this study and the main research issues they address. A literature review with discussion of the research issues is provided in Chapter II. A detailed account of study design is provided in Chapter III. Results are presented in Chapter IV, and conclusions about the applicability of different models of value structure for the examined product category and sample, as well as for possible generalizations, are drawn in Chapter V.

TABLE I

SUMMARY OF VARIATIONS IN EXPERIMENTAL DESIGN

Main Research Issue(s) as Pertaining to Literature (pp. 49)	Study Variation(s)	Explanation
(1) to (4)	Moderately new and complex product category	This is in contrast to many other studies which applied conjoint to familiar product categories, as for instance student apartments, or transportation mode.
	Two (2) administrations of experiment	This allows tests for reliability over time.
(1) and (4)	Systematic variation in the type of attributes comprising the stimulus set(s) (A1, A2)	This variation constitutes a test of reliability over attribute set. It tests effects on the dependent variable purchase likelihood, and on part-worth estimates and importance of attributes
(2) and (4)	Two (different) fractional factorial designs with limited first-order two-way interaction terms (FF1, FF2)	This variation constitutes a test of reliability over stimulus set. It tests effects on the dependent variable purchase likelihood, and on part-worth estimates and importance of attributes
	OLS regression as the estimation procedure	
(2) and (4)	Two types of value structure modeling techniques, traditional conjoint model (TC) and self-explicated model (SE)	Test of (convergent) validity
(3)	Four variations in grouping of respondents for parameter estimation: - individually, - a posteriori; • hierarchical clustering (HIC) • non-hierarchical cluster algorithm (NHC) • fuzzy clustering (FUC)	This tests for conjectured improvements in predictive accuracy when estimating parameters on the basis of customer segments. Fuzzy clustering is a new grouping approach that allows for cluster overlap.

1.6 Delimitations and Limitations

Concerning the many possible extensions to conjoint analysis, the choices were limited to those outlined in Table I on page 23 which seemed most promising in pursuit of the main purposes of this study, in providing answers for the research issues, and in being possible without a budget. In particular, there was no simultaneous data collection for a familiar product category and data collection methods for the new product category were held constant. The number of treatments, the number of respondents, and the lack of a budget required a convenience sample. However, in a large number of studies, this has not been shown to be of influence for the findings (Moore and Semenik 1988; Green, Helsen, and Shandler 1988; Green, Krieger, and Bansal 1988; Johnson, Meyer, and Ghose 1989; Green and Helsen 1989; Moore and Holbrook 1990; Akaah 1988, 1991; Elrod, Louviere, and Davey 1992; Steenkamp and Wittink 1994).

Major limitations of conjoint analysis that are generic to the method of course apply to the current study as well:

- For many applied marketing problems the number of attributes is large to get a realistic context (ten (10) to fifteen (15) and over, versus the five (5) or six (6) often used in academic studies). However, the desire to not contaminate studies with effects unaccounted for, as for instance respondent attention to all relevant attributes, also lead this study to limit the number of attributes to figures deemed appropriate for this experiment (i.e. nine (9) attributes)[1]. Furthermore, for too large a set of factors, responses may be unreliable because of respondent fatigue or simplifying strategies not employed in real decision contexts.

[1] Nine attributes constitute the empirically found upper bound concerning capability of people to process pieces of information simultaneously (7 ± 2; Miller 1956)

- Advanced experimental designs, as for instance the inclusion of interaction terms or high fractionation, often force researchers to sacrifice some flexibility for individual respondent-level analysis, i.e. resort to group-level analysis. Other approaches to the size problem are self-explication models and one-attribute-at-a-time data collection procedures which sacrifice task realism. This study should not have suffered from such limitation as it was designed so that individual-level conjoint analysis is still possible.
- It is acknowledged that using ratings of likelihood of purchase only captures one-choice situations. There is no (explicit) provision for no-purchase or multiple purchase choices. However, this alternative is only relevant when one wants to determine penetration of a market with a new product not competing on the same attributes , i.e. competition between product categories, and if one wants to determine what factors modify the utility function of the customer for final choice (see esp. Sheth, Newman, and Gross 1991a, 1991b for such factors). These deliberations are external to the scope of this study which focuses on tradeoffs among alternatives described on the same attribute set.
- Traditional conjoint analyses do have a "flat" choice structure, i.e. no attribute hierarchies are modeled (for a possibly problematic extension to hierarchical conjoint analysis cp. section 2.4.1 of this study). However, it is conjectured that such interattribute effects, if they exist, are caught with interaction terms in the model.

It seems that a balance is necessary between what is desirable as the conceptual model, and what is feasible from a respondent standpoint (Wyner 1992a and 1992b, p. 46).

1.7 Organization Plan

Chapter II contains a review of the literature that is relevant to this research. Chapter III presents the study design, procedure, research questions, and hypotheses. The results obtained from the study, the answering of the research questions, and the validation of the hypotheses with empirical data are discussed in Chapter IV. Major findings, contributions to marketing practice and theory, limitations, and directions for future research are presented in Chapter V.

CHAPTER II

REVIEW OF THE LITERATURE

This chapter presents a review of the literature on the major conceptual and methodological issues addressed in this study. First, the nature of customer value is examined as a central concept in marketing, and as it pertains to the issues of the current study, mainly the selection of attribute type for the evaluation task of a conjoint experiment. Then, current approaches to measurement of determinant attributes or customer value components, i.e. respective benefit and cost components associated with a product alternative are reviewed, briefly. Third, theoretical bases for conjoint and related measurements are examined, and respective rationales for preferring one over the other shall be provided. Fourth, important research issues and methodological problems in conjoint analysis are addressed, and a rationale for examination of four of them in the current study are provided. Fifth, new suggestions and approaches for improving conjoint measurement with segment-level part-worth estimation are discussed. Specifically, the concept of grouping subjects with fuzzy clustering is introduced. Finally, reliability and validity concepts are reviewed to clarify and justify respective choices of design made in this study.

2.1 Nature of Value

Clarifying the nature of value is important in order to evaluate current measurement concepts of value, i.e. how product features and other benefits or utility derived from the product are translated into value perceptions, and how these, in turn, are translated into money equivalents. One such model, and the one utilized in this study, is depicted in Figure 3 on page 17. Furthermore, knowing the nature of value is important for the question which perceptual dimension, i.e. which types of determinant attributes, should be measured in the evaluative stage of customer decision process. Currently, there is no universally accepted system language for customer value constructs among marketing researchers, and relevant constructs themselves are not agreed upon. Alternative approaches to (perceived) value conceptualizations are

(1) value expressed as a ratio of benefits and prices,

(2) the means-end chain approach, and

(3) the development of generic value taxonomies.

These approaches are not mutually exclusive but rather represent different attempts to conceptualize the translation process from product or service attributes to value perception. They will be used in this study to structure the determinant variables, and provide a rationale for tests of effects of attribute type on part-worth utilities.

2.1.1 Ratio Form of Value

The concept of value is commonly defined by two components — benefits (Zeithaml 1987, proposition V1, p. 21; "performance" in Potter 1988, p. 25) and price ("sacrifice components" in Zeithaml 1987, proposition V2, p. 22). It is usually expressed as the difference between, or the quotient of these two concepts (Kotler 1991, p. 291; Christopher 1982, p. 39; Hauser and Urban 1986, p. 450; Monroe, Rao, and Chapman 1987, p. 204; Haas 1989, p. 365; [added by author]):99

$$\text{Value} = \text{Benefits} - \text{Price} \qquad \text{(E2.1a)}$$

$$\text{[Perceived] Value} = \frac{\text{[Perceived] Benefits}}{\text{[Delivered] Price}} \qquad \text{(E2.1b)}$$

In addition, value is a relative concept. It has meaning only with respect to the proper context, i.e. in reference to benefits absent without the product or service, or compared to some other, competing product or service offerings. It is expressed in monetary units, and high value is equivalent to many benefits or high performance per monetary unit. Therefore, value measurement involves the translation of benefits or performance — more specifically, the customer's perception of these — into money equivalents.

There is, however, no clear concept in the literature as to what "benefits" and "price" encompass. Most marketers will assert that in order to understand customer buying behavior, it is necessary to look at customers' perceptions of the benefits of a product. They also agree it is necessary to consider customers' perceptions of what they must give up, i.e. sacrifice, to obtain a product, including the perceived monetary and nonmonetary price. A third contention and premise is that buyers will buy the product that offers the highest "delivered value" (= value maximization or value priority hypothesis). While Kotler and Christopher define this "delivered value" as the *difference* between "total customer value" and "total customer price" (Kotler 1991, pp. 289; Christopher 1982, p. 39; cp. (E2.1a)), most others cast this concept into the *ratio* form (Monroe Rao, and Chapman 1987, p. 204, and others; cp. (E2.1b)). The problem with these two definitions is that they postulate general applicability over a wide range of products and their respective benefit and cost characteristics, as well as over a diverse population and their respective benefit and cost attributions to these product characteristics (mostly expressed in form of preference judgments). While tests of an additive, respectively subtractive model form of those preferences have found supporting evidence, so have tests for the ratio, respectively multiplicative form (Anderson 1981, pp. 29).

In an exploratory study designed to reveal the definitions and relations between price, perceived quality, and perceived value, Zeithaml (1987, p. 1) grouped consumer opinions of value into four definitions: (1) value is low price; (2) value is whatever I want in a product; (3) value is the quality I get for the price I pay; and (4) value is what I get for what I give (Zeithaml 1987, pp. 18). While these four definitions involve value, price, benefits, and quality, only the last definition is consistent with the conceptualization of value as a difference or ratio of (several) benefit components weighted by their evaluations. The important point, however, is whatever measurement instrument is used to capture value perceptions, it must be able to model these diverse, idiosyncratic forms. Conjoint analysis promises to accomplish this.

Though each consumer definition has its counterpart in the academic or trade literature on the subject, only the latter is able to encompass all four definitions. The first definition reveals the salience of price for specific customer segments, or in specific product comparisons. The second definition is equivalent to economists' definition of utility, i.e. a subjective measure of the usefulness or want satisfaction that results from consumption. It emphasizes benefits derived from consumption. The third definition conceptualizes value as a trade-off between price and quality, while the fourth simply extends the scope of benefits from quality to other possible benefits, and the monetary price components to nonmonetary ones. Using different semantics, the utility-per-dollar measure of value used by Hauser and Urban (1986, p. 447), and others is equivalent to the fourth definition. Finally, all these expressions of value can be captured in this overall definition:

> Perceived value is the consumer´s overall assessment of the u·ility of a product based on perceptions of what is received and what is given.

While what is received varies across consumers (i.e. some may want volume, others high quality, still others convenience) and what is given varies (i.e. some are concerned only with money expended, others with time and effort), value represents the trade-off of the salient give and get components (Zeithaml 1987, pp. 18 - 20). This suggests that not only may model form be highly idiosyncratic, but the type of attributes relevant for evaluation may vary substantially from individual to individual, too. Nevertheless, when we want to measure value components, we must decide in advance about the characteristics, i.e. the determinant attributes, on which product alternatives shall be judged. Thus, accuracy of measurement in terms of customer value hinges on the selection of attributes used for measurement, i.e. the type of attributes (i.e. concrete, physical, product-referent attributes, or abstract, user-referent attributes), the number of attributes included in the description, and the values (i.e. levels) an attribute may assume. There is always the danger that attributes are not included which may be relevant for particular individuals´ evaluations. Furthermore, there is ample evidence that utilities for particular attribute levels vary widely across individuals. Therefore, one may reasonably conjecture that individual-level estimation of value components, expressed as attribute level utilities in a conjoint experiment, may be superior in terms of predictive accuracy to segment-level estimations, where utilities of benefit components are averaged across individuals.

2.1.2 Means-End Chain

The means-end chain approach to understanding cognitive structure of consumers is another approach to conceptualizing value. It holds that individuals retain product information in memory at several levels of abstraction (Young and Feigin 1975, Geistfeld, Sproles, and Badenhop 1977, Myers and Shocker 1981, Olson and Reynolds 1983, Corfman 1991a), ranging from the simplest level of physical product attributes to complex personal values. These values may be the result of judgments made on the basis of cognitive assessment or affect. The central idea is, however, that a product is linked to perceived benefits through a chain of concrete, physical, or measurable product attributes, as for instance MTBF figures (Mean-Time-Between-Failure) for disk drives, their outer measures and weight, and abstract benefit perceptions, as for instance quality, reliability, or serviceability which may theoretically be expressible in concrete product terms but are usually formed through affective cues rather than cognitive judgment. The frame of reference, e.g. prior experience, beliefs, or attitudes, influences the perception of physical product attributes and thus the inferences made about, and summarized in a product´s abstract characteristics, as for instance in perceptions about quality.

The significance of this model for value measurement in connection with conjoint analysis may be expressed in the following questions: Which is the relevant evaluative dimension that should be measured in a conjoint experiment, the physical product attribute (i.e. the one that only refers to the product), or also user-referent attributes (i.e. the ones that refer to general beliefs of the customer) ? Is there a difference in the relative attribution of benefits to attributes and their respective levels depending on the presence of specific types of attributes ? If there are such dependencies, and if they are 'large', this could pose a serious problem for the validity of conjoint measurements (Oppewal, Louviere, and Timmermans 1994, p. 104) as conjoint analysis assumes that the presence of one attribute does not alter the relative benefit attributions to two other variables (Anderson 1981, p. 18; independence assumption). At the very least, 'large' effects would force tests of interactions and inclusion in the conjoint model estimated.

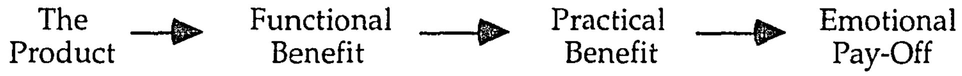


Figure 4. Means-End Chain Model by Young and Feigin (1975), p. 73.

Figure 5 depicts the means-end chain idea as expressed by Young and Feigin (1975, p. 73; proposed earlier by Rokeach 1973). Table II (p. 35), adapted from Zeithaml (1987, p. 7), lists selected means-end chain models and their proposed relationships with quality and value. However, classifications within a particular level as well as between adjacent levels seem arbitrary and artificial. In particular, judgments about them seem highly idiosyncratic due to a lack of common understanding of these terms among individuals. The respective means-end chains in Table II on page 35 are only understandable in light of the specific situations and purposes for which they were developed. The means-end chain concept does not seem suitable to serve as a generic model for deriving perceived customer values. It delineates, however, hierarchies of comparability and evaluation, and gives some guidance for the possible translation process from product attributes to value perceptions, with respective repercussions on the design and conduct of measurement procedures. Specifically, means-end chain models suggest that user-referent attributes are more relevant and more direct measures of customer value than product-referent or technical attributes.

An important question is if the inclusion of more abstract, user-referent attributes, like for instance 'quality', 'convenience', or 'reputation', in a conjoint measurement influences the evaluation of the more physical attributes, and to what extent ? The attractiveness of only minor influences of user-referent attributes on product-referent attributes lies in the possibility to divide up a large number of relevant attributes into (non-overlapping) sets, the values of which can be evaluated in separate experiments without resorting to compromise designs. A detailed account of this problem and its relevance for this study is further provided in section 2.4.

TABLE II

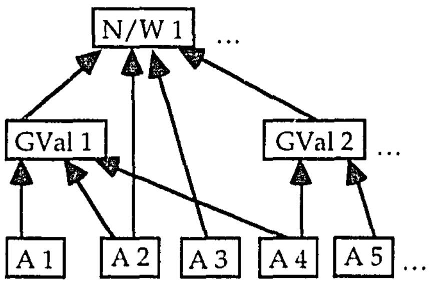
SELECTED MEANS-END CHAIN MODELS AND THEIR PROPOSED RELATIONSHIP WITH QUALITY AND VALUE

Scheme	Attribute Level	Quality Level	Value Level	Personal Value Level
Young a. Feigin (1975, p. 73)	Functional benefit	Practical benefit	Emotional Payoff	
Geistfeld, Sproles, and Badenhop (1977)	Concrete, unidimensional, and measurable attributes (C)	Somewhat abstract, multi-dimensional, but measurable (B)	Abstract, multidimensional, and difficult to measure attributes (A)	
Cohen (1979)	Defining attributes	Instrumental attributes		Highly-valued states
Myers and Shocker (1981)	Physical characteristics	Pseudo-physical characteristics	Task or outcome referent	User referent
Olson and Reynolds (1983) similar to Rokeach (1973)	Concrete attributes	Abstract attributes	Functional consequences, psychosocial consequences, instrumental values	Terminal values
Corfman (1991a, p. 370)	Concrete attributes or dimensions, features	Abstract attributes or values, basic values	micro- and macrofunction	Overall utility or worth

2.1.3 Generic Value Taxonomies

Finally, recognizing shortcomings in current business literature concerning taxonomies for perceived value derived from hierarchical and nonhierarchical benefit constructs, led to the attempt to devise generic value taxonomies applicable in a wide variety of customer choice situations for consumer and/or industrial goods. The fundamental premise of these taxonomies is that market choice is a multidimensional phenomenon involving multiple, primarily independent 'values' which denote the constructs or the class of constructs used for evaluation of product alternatives. Table III (p. 37) tabulates generic taxonomies of benefits and sacrifices (mostly termed "values" and "costs" or "prices" in their original references). The model behind these taxonomies is depicted in Figure 5.

Their significance for conjoint measurement is that these taxonomies provide a means to check for possibly omitted dimensions in devising the attribute set for a product or service alternative to be evaluated. However, conjoint analysis seems to be more flexible as a measurement tool, insofar as any item evaluated by potential customers may have benefit or cost character, depending on its relationship to all other attributes on which a product alternative is evaluated. This flexibility may also be viewed as a partial relaxation of the assumption of independence from irrelevant attributes.



Level of Variable

Satisfaction of need or want

Generic values or benefits (higher-level abstractions)

Specific choice attributes (e.g. physical appearance or performance characteristics, aesthetics, etc.)

Effects are only allowed from lower to higher levels of abstraction (need not be the next higher one)

Figure 5. General Model of Value Taxonomies (Levels of Abstraction).

TABLE III

SELECTED GENERIC VALUE TAXONOMIES

Author	Benefits	Price (Sacrifice)	Value Terminology	Applicability
Sheth, Newman, a. Gross 1991a, p. 160; Sheth et al. 1991b, pp. 7	Functional value Social value Emotional value Epistemic value Conditional value	Money Time price Allocation of effort	Consumption values	Consumer marketing
Kotler 1991, pp. 290	Total customer value = Product value Services value Personnel value Image value	Total customer price = Monetary price Time cost Energy cost Psychic cost	Delivered value	For consumers and industrial buyers
Monroe, Rao, and Chapman 1987, pp. 204	Relative use, exchange, and aesthetics = Physical attributes Service attributes Technical support relative to particular use of product	Cost = Purchase price Acquisition cost Transportation Installation Order handling Risk of failure	Total relative value	For consumers and industrial buyers
Forbis and Mehta 1981, pp. 34	Physical features of product (functions, added application, technical reliability) Other attributes (intangibles, e.g. delivery reliability, service responsiveness, even brand name, satisfaction of personal or social needs)	Life cycle cost = Purchase price Start-up costs Post-purchase costs (maintenance and operations)	Economic value to the customer (EVC)	Industrial buyers

2.1.4 Conclusions About the Nature of Value and Value Perception

Objective value and perceived value are not equivalent, i.e. the same function performed or the same physical attribute of a product may lead to vastly different value perceptions dependent on the user or the intended use of the product. Perceptions of benefits may be concrete (e.g. color, weight, height) or abstract (e.g. sturdy, robust, reliable, flimsy) evaluations or judgments formed from intrinsic attributes of the product (e.g. its physical or performance characteristics) and extrinsic attributes that are not part of the actual physical product (e.g. price, brand name, packaging, warranty). The benefit components of value include salient intrinsic attributes, extrinsic attributes, and relevant higher-level abstractions, as for instance perceived quality or convenience (Zeithaml 1987 and 1988). The significance of conjoint measurement with regard to diverse types of attributes for value perception, especially product-referent and user-referent ones, lies in its potential to measure these attributes′ respective relevance for value judgments in a common space, denoting their relative impact on customer evaluations of product alternatives. It allows for the proverbial comparison of apples and oranges.

2.2 Current Measurement Approaches

Value measurement is regularly being discussed in connection with pricing, and to a lesser extent with new product development. A plethora of empirically validated quantitative models is available in consumer marketing, especially in the field of attitude research. Lilien, Kotler, and Moorthy (1992, p. 31) contend, ideally, a model of buying behavior, and specifically value measurement, would

- identify and measure all major variables making up a behavioral system,

- specify fundamental relationships among the variables,
- specify exact sequences and cause and effect relationships, and
- permit sensitivity analysis in order to explore the impact of changes in the major variables.

However, for the sake of parsimony, most consumer behavior models only attempt to do a portion of this job (Lilien, Kotler, and Moorthy 1992, p. 31), and even then they get extremely complicated and need large sample sizes to test for predictive validity. It does not seem feasible for virtually any study to accommodate the extensive modeling requirements suggested by Lilien, Kotler, and Moorthy (1992, p. 31) to achieve predictive validity of choice behavior. Therefore, most commercial applications measuring attributions of utility to product profiles do not derive them with elaborate measurement procedures, but use self-explicated utility measures. A national marketing research firm[2] developed a questioning tool based on self-explicated preferences that can be administered via telephone (CASEMAP; Srinivasan 1988, Srinivasan and Wyner 1989).

With self-explicated utility measurement, subjects perform two tasks:
First, they are asked to rate desirability of attribute levels (on a 0 to 100 point interval scale for each set of attribute levels), or are asked to distribute e.g. 100 points over respective levels of an attribute that indicate within attribute desirability of levels. Second, respondents are asked to rate, or again distribute another number of points according to the importance of specific product attributes. After normalization, the utility of an alternative is simply the sum over all products of level desirability times importance of the attribute (see equation E3.1.3 on page 81). In comparisons with conjoint models, this provides an easy (convergent) validity check for the conjoint

[2] M/A/R/S developed an automated questioning tool based on Srinivasan's CASEMAP program.

model, and sometimes yields equivalent accuracy (Leigh, MacKay, and Summers 1984; Srinivasan 1988; Green and Helsen 1989; Green and Schaffer 1991).

2.3 Theoretical Bases for Conjoint and Related Measurements

Conjoint analysis (Green and Srinivasan 1978) or conjoint measurement (Green and Wind 1975) is a family of methods to measure perceptual and judgmental concepts on the individual level in categorical and metric form. It includes any technique used to estimate attribute utilities based on subjects´ responses to combinations of multiple decision attributes. Conjoint analysis, especially its metric form, is based on information integration theory (IIT) as first summarized in two books by Anderson (1981 and 1982), and developed by him and many other researchers before him (Bettman, Capon, and Lutz 1975; Louviere 1974; Slovic and Lichtenstein 1968; Slovic, Fischhoff, and Lichtenstein 1977). In contrast to aforementioned approaches to value measurement it has a theoretical and empirical basis in psychology, and, if measurements are repeated, has an error theory to allow for statistical tests of alternative models of customer value structure as the immediate basis for their decision making. This allows for greater scientific rigor in empirical estimations of conjectured value structures. It is based on four intimately related concepts: stimulus integration, stimulus valuation, cognitive algebra, and functional measurement (Anderson 1981, p. 2).

Stimulus integration is a central concept in establishing the link between thought and behavior. Both, thought and behavior, are influenced by the joint action of multiple stimuli, rendering modeling with assumptions of multiple causation a necessity for understanding or prediction. IIT is interested in two questions with respect to stimulus integration: (1) Given effective stimuli, how are they combined or integrated to produce the response (compositional perspective or synthesis) ? In terms of conjoint analysis this is a question of model form, i.e. additive, multiplicative (i.e. interactive), and mixed forms, and in an analogous manner it involves subtractive or ratio form of customer value. (2) Given the response, what were the effective stimuli (decompositional perspective or analysis) ? The second question constitutes merely a different perspective of the first, but is often the more important question when applying a measurement instrument due to the necessity to reduce information overload and guard against respondent fatigue.

Stimulus valuation is commonly distinguished as occuring at two levels: (1) At the physical level where stimuli are observable, measurable, and potentially controllable in an experimental setting. However, these are distant, indirect, and partial causes of thought and behavior. (2) At the psychological level where (psychological) stimuli are the immediate causes of thought and behavior. The translation process or chain of processing from physical stimulus into its psychological counterpart is represented by the valuation operation and modeled mathematically with respective variables and connective operators. IIT stresses the particularity of individual valuation processes and according structure, and conjoint analysis provides a methodology to measure individual differences in stimulus valuation. The only problem with this approach to value measurement are influences from attitudes and prior beliefs which may be activated not by physical (product) cues but by other psychological constructs.

Cognitive algebra is the term used for the empirical fact that stimulus integration frequently obeys simple algebraic rules and is a reasonably good, high-level approximation of actual subject processing. In the absence of more detailed knowledge about the brain and body´s functioning in stimulus processing, the human organism often at least appears to be averaging, subtracting, or multiplying stimulus information to arrive at a response. An important question in this research study is if segmentation of subjects′ responses to stimuli and subsequent re-estimation may be able to improve over biased individual-level estimates.

The concept of functional measurement includes two aspects: (1) It is possible and appropriate to represent stimuli numerically, which is also implicit in the notion of cognitive algebra. (2) Even if stimuli at the physical level cannot be described in numerical terms, measurement of psychological stimuli can be accomplished with algebraic descriptions of stimulus processing as revealed by according responses. Functional measurement then simply denotes that the algebraic rule "functions" to explain the response. It may therefore also be termed "processing function fitting" or "processing function approximation."

Finally, Anderson makes it explicit that this view of the individual organism as an integrator of stimulus information with judgments exhibiting specific algebraic properties is part of the "Zeitgeist" (Anderson 1981, p. 3). Implicit in this statement is acknowledgment that IIT, or its primary method conjoint analysis, may very well be augmented or replaced by a better paradigm for value judgments, if such becomes available (which is in welcome contrast to frequent history of changes in scientific paradigms; cp. Kuhn 1970, Kosko 1993). Several such competing approaches to (evaluative) preference and choice modeling are briefly discussed next.

One such approach are attitudinal models. As this approach has been dealt with extensively in the marketing literature (see Sheth, Newman, and Gross 1991a, 1991b; Engel, Blackwell, and Miniard 1990; Ajzen and Fishbein 1980), only one brief description of a recent comparative study with conjoint analysis by Nataraajan (1993) may be furnished. This study examined one utilitarian and one attitudinal approach to modeling of value judgments; traditional conjoint analysis and the theory of reasoned action (TRA). TRA conjectures intention as the immediate antecedent of behavior. It is a variation on the extended Fishbein model (Fishbein 1963; Fishbein and Ajzen 1975; Ajzen and Fishbein 1980; for a test of this model's convergent, discriminant, and predictive validity see Burnkrant and Page 1982, and Bagozzi 1982). Intention has two components: (personal) attitude and (social) norms. The relationships of the model may be expressed in equation E2.3.1:

$$B \sim BI = w_1 \cdot A_B + w_2 \cdot SN \qquad \text{(E2.3.1)}$$

where B is overt behavior (~ means "approximately corresponds with"),

BI is behavioral intention (subjective probability of intending to perform behavior B),

A_B is attitude toward performing behavior B (e.g. attitude toward buying a brand; note that this is not attitude toward the brand itself),

SN is subjective norm (normative influence; the collective perceived influence from "important others"), and

w_1 and w_2 are empirically determined weights denoting the relative influence of the two components.

A_B is determined as ($\sum_{j=1}^{N} b_i e_i$) where b_i is the subjective probability that performing the behavior will result in outcome i, e_i is the individual's evaluation of outcome i, and n is the number of salient outcomes. SN is determined as ($\sum_{j=1}^{N}$ $NB_j MC_j$) where NB_j is the belief that referent j thinks the individual should/should not perform the behavior, MC_j is the individual's motivation to comply with referent j, and N is the number of salient referents.

Nataraajan (1993) tests both model forms of evaluative choice in an experimental setting and concludes that for the product class studied, conjoint analysis has a significantly higher first choice hit rate. However, he cautions that this may be due to the peculiarity of the product class utilized for comparison and may not be generalized over different product classes (Nataraajan 1993, p. 378). It may further be added that study objective (here, prediction and not explanation was of main interest) as well as other problem, method, and procedural contexts may be responsible for the outcome. However, there, as in other contexts of value structure modeling, conjoint procedures have two appealing advantages over attitudinal models: (1) They allow estimation of model parameters for and on the basis of only one individual respondent while attitudinal models usually use one or the other form of aggregated parameter estimation over at least a subsample (in Nataraajan´s case the estimation of brand specific beta weights). (2) The decompositional approach of conjoint analysis relying on statistical derivation of the components (part-worths) of customers´ overall judgments seems to be better than a compositional approach relying on direct customer input. Decision makers (not only in connection with consumption decisions) were found to often not be able to reveal estimations of values for their decision components (for additional citations of studies in the sixties see Green and Schaffer 1991, p. 476).

Another utilitarian approach to value measurement from decision theory is Saaty´s Analytical Hierarchy Process (AHP; Saaty 1980; for critical remarks see Dyer 1990; for an analysis of connections between hierarchies, objectives, and fuzzy sets see Saaty 1978). Like other MAUT methods it typically focuses on small numbers of decision makers involved in high-level decisions. In contrast to other MAUT methods, though, it is more descriptive of the decision process than normative. It also typically involves in-depth questioning, a time-consuming and highly involved data collection method. But, with consumer decisions, time and cost considerations as well as unclear effectiveness of such procedures in consumer contexts usually forbid high demands on individual response capability. Considering this aspect, conjoint analysis allows estimation of individual value structure when at the same time putting minimal burden on respondents´ task capability. This property of conjoint analysis favors it over the AHP procedure.

A highly favored approach to preference modeling in recent years in marketing is based on factor analytic and structural equation models. In these models there are several physical and/or psychological states, and with the latter approach it is possible to test the relationships between those states and a number of internal (psychological) factors conjectured to constitute direct antecedents of behavior. The two main drawbacks of this approach are: (1) There is no direct connection from attribute to outcome. (2) Estimation procedures of these models are not applicable for individual, only for aggregate analysis. This latter problem may not be critical for predictive purposes. However, the lack of individual analyses makes important information about individual differences concerning product evaluation and according indications of profitable business opportunities inaccessible for marketing decisions, specifically information about benefit segmentation, design preferences, and appropriate tactical and strategic decisions. While for instance Sheth, Newman, and Gross´ (1991a and 1991b) procedure to measure "consumption values" indicates the relative influence of each of five generic factors for choice, this procedure´s data does not allow for the incorporation of attributes other than suggested by the measurement procedure but that may nevertheless be desirable to know from a managerial perspective; and individual data of Sheth, Newman, and Gross´ procedure does not allow for analyses other than for the prespecified objective of revealing contributions of five factors for market choice. Segmentation, for instance, has to be done on the basis of a priori defined control variables. Conjoint analysis, in turn, allows for a much broader array of objectives and analytical procedures.

Finally, the process of decision making based on value perceptions and operationalized with conjoint analysis shall be explained with Figure 3 on page 17. The top row of labels shows assumed mappings from physical reality to judgments about product preferences and choices. The bottom row of labels shows inputs and outputs, i.e. the static components of respective mappings. The process of value perception is conjectured to begin with psychophysical judgments about physical reality, resulting in perceptions or beliefs about positions or levels of a stimulus on a number of (physical) attributes (mapping f_1).

For instance, if "convenience" is a determinant decision attribute for shopping centers, consumers might consider physical factors, such as travel time, parking costs, parking space, hours of operation, parcel carryout, acceptance of credit cards, and the number or type of other services, facilities or offices (e.g. banks, post office, library, travel agents, etc.) to form impressions of "convenience" of a particular shopping center (alternative n in Figure 3´s notation). However, customers may not perceive physical variables, such as travel time or the amount of parking space, in physical terms, but rather use physical cues to make psychophysical (i.e. perceptual) judgments (Mehrotra and Palmer 1985, p. 84; Myers and Shocker 1981, p. 225) and then remember and use for judgments only the abstract, psychological construct of "convenience" (i.e. the amount of "convenience" a particular shopping center is believed to possess). Some of the physical variables may be perceived more accurately and used directly for judgments, while others may only be used to evoke the psychological constructs which they contribute to. Some beliefs about variables may even be based more on prior beliefs about for instance a brand name than on physical cues from the (product) stimulus. In any case, this process of forming impressions and beliefs about positions of various choice alternatives on decision attributes (i.e. determinant choice criteria) involves the integration of perceptual information (mapping f_1).

Having formed these perceptions, consumers make (personal) value judgments about how good or bad it is for an alternative to be positioned in a particular way on each attribute (mapping f_2). The result of this evaluation process is an attitude or utility for each attribute, i.e. $V(S_{nj})$. By combining attribute valuations in some way (which may be modeled algebraically), consumers arrive at an overall evaluation, U(n), for each brand (i.e. decision alternative), illustrated with mapping f_3. This evaluation process can be inferred from overall judgments about alternatives by assuming (and sometimes testing for) ways in which consumers combine (i.e. integrate) information about different determinant attributes to arrive at an overall evaluation for each choice alternative. It is exactly this integrating process of combining attribute information that is modeled with conjoint analysis techniques as the methods to elicit information integration behavior. Conjoint analysis permits to study these cognitive processes and develop statistical approximations to them by specifying the integration model (additive, multiplicative) and estimating part-worth utilities, i.e. attributions of benefits to decision criteria.

While mapping f_3 results in an overall evaluation about how good or bad each alternative is judged to be (i.e. U(n), the alternatives´ utilities), final choice decisions are contingent upon factors that are independent of the choice set and their respective attributes, as for instance the available budget, urgency of need or want, availability of an alternative at specific locations, inclination to comply with judgments of referent others, and so on. Therefore, probability of purchase (i.e. purchase likelihood, P(n)) may be different from an individual´s overall evaluations and modeled separately with mapping f_4.

While in principle observations may be made at each step of this translation process, i.e. for each input/output pair in Figure 3 on page 17, most conjoint studies to date involve experiments with physical attributes as predictor ("independent") variables and likelihood of purchase as the criterion variable. The respective functional form of conjoint model is then responsible for capturing respondents´ information integration process. Empirical determination of this process usually spans all four mappings which may be summarized algebraically in equation E2.3.2 by elementary substitution of terms:

$$P(n) = p(n|A) = f_4 \, [\, f_3 \, (\, f_{2j} \, [\, f_{1j} \, (X_{nj}) \,] \,) \,] \qquad \text{(E2.3.2)}$$

where p(n|A) is the probability of selecting the n-th stimulus from choice set A of n (product) alternatives; usually, a direct surrogate measure, likelihood of purchase, is utilized to obtain this choice probability.

All four mappings or relationships are assumed to operate in decision making, but commonly only the end points are measured quantities, i.e. determined empirically. It is conjoint measurement´s characteristic to estimate attribute values $V(S_{nj})$, i.e. part-worth utilities, from overall responses to a number of constructed or actual stimuli. Estimation procedures and functional forms of conjoint models are detailed in section one of Chapter III.

2.4 Methodological Problems in Conjoint Analysis

Conjoint analysis has been a prolific area of academic research into value measurement since its introduction into marketing by Green and Rao (1971), and in recent years enjoys a fast-growing number of commercial applications (Wittink and Cattin 1989; Wittink, Vriens, and Burhenne 1994). Since then, some methodological issues have been settled, as for instance the interpretation of ratings of product profiles on category-rating scales as interval-valued versus early contentions such ratings may only be regarded as ordinal, but a number of unresolved issues remain. Discussion of problem areas in conjoint analysis as they pertain to this study may be organized into three (3) phases, and in the rough order in which decisions about the conduct of conjoint experiments have to be made (as outlined in the introduction; p. 17):

(1) Characteristics of the attribute set.
(2) Design of a conjoint experiment.
(3) Segmentation of respondents according to benefits sought.

Early research concentrated on data collection method (i.e. data gathering procedure and type of dependent variable), model form (i.e. additive, interactive), and estimation techniques (type of regression and ANOVA). Recent reviews (Green and Srinivasan 1990; Wittink, Vriens, and Burhenne 1994) and examination of the literature suggests deficiencies in the following areas:

(1) There has been a lack of examination of the relationship between type, number, and levels of attributes used for evaluation and the resulting value structure, i.e. how characteristics of the attribute set influence resulting part-worths and importance of attributes. Investigators have limited their research primarily to

the study of familiar product categories and have tended to focus their attention on technical aspects of conjoint method.

(2) There is a continued debate as to whether customers´ value structure is sufficiently modeled with an additive model or if interactions between attribute (levels) are necessary to adequately capture attribute value perceptions (i.e. questions of model form).

(3) There is ongoing effort under way to improve predictive accuracy by grouping respondents into segments and estimating part-worths for respective segments. This is currently an area of intensive research, and is also the focus of this study. According details are provided in section 2.5 of this chapter.

(4) It is not clear within the research community what tests and testing procedures establish reliability and validity in a conjoint study. Additionally, there is disagreement over the appropriate measures to use. This lack of agreement concerning methodological concepts threatens the usefulness of past research. This issue is dealt with in section 2.6.

These shortcomings and ambiguities in the literature have important theoretical and applied significance. They formed the basis of this investigation.

2.4.1 Attribute Set

The first research problem, a lack of examination of the relationship between type, number, and levels of attributes used for evaluation of products and services may be due in part to a preoccupation with developing a proper model form for representing customer decision structure. In putting together a value measurement model and designing a respective study the issue of attribute type is dealt with in practical terms. Focus groups, in-depth customer interviews, internal corporate expertise, and trade literature are some of the sources used for structuring the set of attributes and levels guiding the rest of the study. However, the managerial desire to choose product attributes that are actionable in terms of the marketing mix may not be representational of the evaluative constructs and processes used by customers, as has already been outlined in section 2.1 of this chapter. The types of attribute used to evaluate products may be classified into product-referent and user-referent attributes (Myers and Shocker 1981). Product-referent attributes mainly denote physical characteristics of the product, as for instance its weight or size, but also non-physical characteristics, as for instance a warranty that comes with the product. User-referent attributes denote prior beliefs, abstract, and multidimensional constructs, as for instance reputation, quality, or convenience. While there is nothing wrong in choosing only physical attributes for evaluation, its impact on resulting part-worth utilities is as yet unknown. It may well be conjectured that the type of attribute included in the experiment is an important source for variability. If this is true, and to what extent, however, is unknown.

The selection of attributes is influenced by two deliberations:

1. Relevancy for customer evaluation.
2. Relevancy for business objectives, in particular if attributes are actionable for the product manager.

Relevancy for customer evaluation may strongly suggest the inclusion of user-referent attributes, as for instance quality, or reputation and importance of a brand name. However, such attributes are not as easily acted upon as on physical attributes, like for instance the size of keyboard for a computer. If, as in product design, the marketer wants to improve his product´s position on this attribute, he needs to know what other characteristics influence quality or reputation, which necessitates an additional step in measurement. Oppewal, Louviere, and Timmermans (1994) conducted a multistage experiment of store image using a hierarchical structure of attributes: user-referent attributes like convenience or appearance (they term them "general" attributes) are described with subsets of product-referent attributes. One particular store profile is described by the attributes of one of these subsets and the other, user-referent attributes. This does not only keep the evaluation task for one conjoint experiment manageable by limiting the number of attributes per profile, but it also keeps a user-referent attribute actionable if its part-worth utility and importance suggests action to improve the product´s position on this attribute.

While this approach has some advantages over so-called bridging designs for large numbers of attributes, it also has two shortcomings:
First, several conjoint experiments are necessary to measure value structure. Second, their selection of subsets for the user-referent attributes may not capture the whole extent of items that are determinant for those attributes´ part-worths, hence distorting the true utility for these attributes. Oppewal, Louviere, and Timmermans (1994, p. 101) report their results only partially support the hierarchical structure and predictive validity. The direction of these distortions across experiments was not equivocal. They conjecture that context dependency (through introduction of user-referent attributes) could be a larger problem in conjoint experiments than commonly assumed (p. 104). Thus, it may well be conjectured further, that the inclusion of user-referent attributes in a conjoint task has negative effects on part-worth stability with resulting negative effects on predictive accuracy. However, these results are obtained for

aggregate estimations, not for invidual-level part-worth estimates. In a study by Gensch and Ghose (1992) which examined the effect of the type of independent variable included in the conjoint task for individual-level choice models, attributes versus underlying latent dimensions (factor scores), found inconclusive results (p. 36). Attribute evaluations showed higher predictive validity for homogeneous populations (i.e. segments), and evaluations of latent dimensions showed higher predictive validity for heterogeneous populations. Green and Srinivasan's (1978) suggestion to construct "superattributes" (for highly correlated ones) does seem to create similar problems for estimation. Therefore, this study tests the effects of attribute sets with and without user-referent attributes (sets A2 and A1, respectively).

Other recent studies found additional effects pertaining to the attribute set:

Moore and Semenik (1988) tested the impact of different numbers of attributes (five, eight, and twelve) in the master design and generally found a substantial decrease in predictive validity from a design with eight (8) attributes to a design with twelve (12) attributes (Moore and Semenik 1988, p. 269). This may be regarded as confirmation of the conjecture that nine (9) attributes used in this study constitute an upper bound for full profile conjoint experiments.

In a simulation study, Darmon and Rouziès (1989) examined the effect of different attribute level spacings on conjoint estimates, given a specific curvature of part-worth utilities for a particular attribute. They conclude that attribute levels should be unevenly spaced when there is prior knowledge as to the level utilities' curvature (p. 42). However, in the absence of such prior knowledge, even spacings of the levels seems to recover true part-worth utilities best, on average, while still allowing for detection of nonlinearities (Green and Srinivasan 1978). Therefore, evenly spaced levels were used for continuous attributes in this study.

A most recent study by Steenkamp and Wittink (1994) confirmed a number of levels effect on importance of attributes (Wittink, Krishnamurthi, and Reibstein 1989), i.e. the more levels an attribute has in the profile description in comparison with other attributes in the study, the higher its importance. The number of levels effect was on average less than ten percentage points of part-worth utilities for differences in levels of four (4) versus two (2). Thus, the combination of two (2) and three (3) levels for respective attributes in this study may not be considered of substantial influence on part-worth estimates.

Another potential threat to predictive validity of conjoint analysis studies pertaining to the attribute set are correlations between attributes and nonrepresentative designs. It is well known that correlations between predictors of linear models distort estimation of parameters. But much applied work in conjoint analysis involves an important assumption: cognitive processes underlying evaluative and choice behavior may be complex, contingent, and noncompensatory, but they are often modeled well by simple linear compensatory models (Green and Srinivasan 1978). A study by Johnson, Meyer, and Ghose (1989), however, cautions to differentiate: While positive correlations did not exhibit a sharp decline in predictive validity, negative correlations had predictive validity drop to chance levels (they used a level of 33% for both positive and negative correlations). Therefore, with negative correlations present among attributes, estimation of interaction terms should be included in the model. This study uses a fractional factorial design that allows for estimation of selected interaction terms, in particular for the one negative correlation revealed by the pretest (cp. Appendix I).

Orthogonal designs usually applied in conjoint experiments may create non-representative or unbelievable combinations of attribute levels which, in turn, could distort estimations of value structure. However, studies by Moore and Holbrook (1990, p. 496) and Mehta, Moore, and Pavia (1992, pp. 474 and 475) did not find unbelievable attribute combinations to be of significant effect on predictive validity. The latter study, however, found that removal of unacceptable levels may need approximately 30% fewer paired comparisons if their procedure is applied for Adaptive Conjoint Analysis (ACA), a specific presentation and estimation method for conjoint experiments. Steckel, DeSarbo, and Mahajan (1991) developed a procedure for the creation of acceptable conjoint analysis experimental designs. However, as their designs are not necessarily orthogonal, these designs may even perform worse than orthogonal designs allowing unbelievable attribute level combinations. The effect of cultural environment, however, has been found to be important for estimation of value structure, which is intuitively plausible (Sriram and Foreman 1993, p. 62). In light of current research, this study uses an orthogonal fractional factorial design that does not guard against unbelievable attribute level combinations, but which may pose fewer problems than anticipated, anyways. Table IV on page 56 depicts selected conjoint studies examining effects of attribute characteristics on preference and choice behavior.

TABLE IV

SELECTED CONJOINT STUDIES EXAMINING EFFECTS OF ATTRIBUTE CHARACTERISTICS ON PREFERENCE OR CHOICE BEHAVIOR

Source	Date	Topic
Moore and Semenik	1988	Hybrid ConjA and the impact of a different number of attributes in the master design
Boecker and Schweikl	1988	Individualized relevant attribute sets, not only individual-level estimation
Darmon and Rouziès	1989	Effect of various continuous attribute level spacings
Johnson, Meyer, and Ghose	1989	Linear compensatory choice models fail in negatively correlated environments
Wittink, Krishnamurthi, and Reibstein	1989	Effect of various continuous attribute level spacings for ratings response data
Moore and Holbrook	1990	Non-representative designs (environmental correlation of attributes) resulting from orthogonal arrays seem not to be much of a problem
Steckel, DeSarbo, and Mahajan	1991	Creation of modified fractional factorial designs which are as orthogonal as possible while eliminating unacceptable level combinations; has not been applied yet
Mehta, Moore, and Pavia	1992	Examination of the use of unacceptable levels in ConjA yielded no negative effects on prediction but their elimination necessitates fewer comparisons
Gensch and Ghose	1992	Actual product attributes may be better predictors of disaggregate choice models than underlying latent dimensions (factors)
Oppewal, Louviere, and Timmermans	1994	Hierarchical structure and predictive validity of user-referent attributes is only partially supported
Steenkamp and Wittink	1994	Metric quality of full-profile judgments and the number-of-attribute-levels effect

2.4.2 Model Form and Fractional Factorial Design

The second research problem is continued debate as to whether customers´ value structure is sufficiently modeled with an additive model or if interactions between attribute (levels) are necessary to adequately capture attribute value perceptions. While it is acknowledged that, theoretically, all possible interactions should be included, time and cost constraints often preclude their consideration in designing the study. The inclusion of interaction terms allows for tests of attribute interactions and possibly higher predictive accuracy. However, these come at the cost of increased data collection efforts and decreased parameter stability. Therefore, especially in commercial conjoint studies, interaction effects are sought to be avoided, with their effect on part-worth utilities being largely unknown. Additionally, attributes or combinations are sometimes changed into "superattributes" to avoid the inclusion of interaction terms, which contributes to the first research problem. This suggests at least screening for interactions and conduct of a pretest in order to include them into the design, and check for effects after conduct of the experiment. This is the approach taken in this study.

Realistic decision contexts for a variety of consumers and a variety of products often necessitates inclusion of more than five or six attributes at more than two levels each. Therefore, highly fractional factorial designs are the only feasible method to estimate part-worth utilities for all attribute levels, on an individual basis, and using full profile presentations. A balance is needed between comprehensiveness of evaluative items and parsimony in data collection and model form. While this is not unique to conjoint analysis, conclusions on the basis of Monté Carlo studies, though useful, cannot replace empirical experiments with "real" subjects. Assumptions in constructing the Monté Carlo study, as for instance normality of error term distribution, may not be present in real evaluative situations, and there is often no way to test for the presence of these assumptions. Accordingly, there has generally been a negligence of necessary tradeoffs between parameter reliability and degrees of freedom (DFs). Tests of significance on the individual level are of limited value as ratios of sample size (of the number of profiles) to the number of parameters regularly do not approach higher values than a ratio of 2:1. The main effects models in this study have ratios of 27:16, and the conjoint models with interaction terms have ratios of 27:18. Increasing sample size of respondents obviously does not contribute to increased reliability of parameter estimates of individual-level conjoint models. Testing part-worth utilities by averaging replications (Louviere 1988) confounds effects of reliability over time with effects due to the fractional factorial design. Thus, empirical studies concerning effects of fractional factorial designs on predictive validity is an urgent need.

In general, studies that included variations in the factorial profile did not attribute effects to this methodological variation but to effects from other methodological choices of their studies, specifically to the type of model estimated (Akaah and Korgaonkar 1983; Akaah 1991; Green, Krieger, and Schaffer 1993b). A study that explicitly tested internal validity under alternative profile presentations and under specific environmental correlations of the attribute sets (Green, Helsen, and Shandler 1988, p. 396) did not indicate that part-worths calibrated in the "wrong" environment predict a holdout sample worse than those calibrated in the "correct" environment, which partly contradicts Johnson, Meyer, and Ghose´s (1989) findings for negative attribute correlations. Darmon and Rouziès (1991), however, testing internal validity of conjoint estimated attribute importance weights, found substantial weight distortions, especially under fractional factorial designs (p. 320). Considering the sparse knowledge and contradictory evidence in the literature about effects of fractional factorial designs on estimation of value structure, and considering approaches to improve predictive validity of conjoint analysis with subject grouping (i.e. segmentation) methods, it seems highly desirable to examine effects of fractional factorial designs on part-worth utility estimates. Therefore, this study tests for effects of fractional factorial designs on predictive accuracy with two different factorial sets (FF1 and FF2).

2.5 Respondent Grouping and Fuzzy Clustering

A third research issue, and the focus of this study, is ongoing effort to improve predictive accuracy by grouping respondents into segments and estimating part-worths for respective segments. Usually, part-worth utilities are estimated for each individual. The rationale behind this is the idea that individuals are so idiosyncratic in their value structure that individual estimations should yield highest predictive accuracy, individually and if grouped together. In fact, the capability to estimate individual-level preference and choice behavior instead of resorting to aggregates has been the impetus to use conjoint analysis for marketing purposes, in the first place. Additionally, individual-level estimation allows for examination of value structure useful for marketing objectives other than prediction of market shares or choice behavior, as for instance for benefit segmentation or strategic planning. However, due to the small number of observations with respect to the number of parameters estimated for individual-level analyses, part-worths are very sensitive to variations in the ratings.

There are basically two approaches to improve predictive accuracy, further:

1. Continue to model individual differences more accurately.
2. Group individuals together into homogeneous groups and estimate value structure of these groups.

As to the first approach this means, not only are part-worths estimated for each individual on a set of attributes, but the set of attributes on which a product is evaluated is individualized, too. Boecker and Schweikl (1988) developed a computer program that allows individualization of the relevant attribute set, not merely of estimation method. This procedure´s predictive performance on R^2 and first-choice hit rate (First-Hit) was tested against a traditional individual-level conjoint experiment including the five, on average, most important attributes. Using VCRs as the product and 24 attributes in the master design, Boecker and Schweikl´s (1988) procedure significantly outperformed individual-level conjoint on First-Hit, and outperformed the traditional approach slightly on R^2. This indicates, that even more individualized procedures may improve predictive validity, especially when First-Hit is the performance measure. The caveat, however, is that this result has been achieved and tested on only one type of product, yet, and it came at higher time and cost requirements for conducting the experiment. In particular it demands computer questioning and individual interviews. Thus, the procedure suffers from the same setback as traditional decision analysis: For "mass" evaluation of value structure in a commercially viable setting, marketing managers need procedures that stay within reasonable cost constraints and demands put on respondents. Therefore, the opposite route is taken in this study to improve on conjoint analysis´ predictive validity.

As to the second approach, there are primarily two motivations for respondent grouping in connection with conjoint studies:

(1) To perform segmentation on the basis of benefits sought in order to aid in effective targeting and positioning strategies.

(2) As a way to increase reliability and (internal) validity of conjoint measurement by trading high variance in respondents′ part-worth estimates for increased parameter stability (i.e. less bias in part-worth estimation) when compared to the individual-level approach, which suffers from less variance in respondents but increased bias in part-worth estimates (Hagerty 1985, 1986; Hagerty and Srinivasan 1991, p. 77; van der Lans and Heiser 1992, p. 327).

The first motivation is due to the fact that modern marketing in industrialized countries cannot do without segmentation of the market of its potential customers. Identification of segments critically depends on both

- segmentation base, and
- segmentation method.

Benefits are among the most powerful bases for segmentation (Wind 1978; Urban, Hauser, and Dholakia 1987; Kamakura 1988), and their expression as part-worth utilities derived from evaluation of product profiles with conjoint analysis may be the most popular method for benefit assessment (Green and Srinivasan 1978, 1990). In the US, one third of purposes for conduct of conjoint experiments comprised segmentation (Wittink and Cattin 1989, p. 92: 1. new product/concept identification 47%, 2. competitive analysis 40%, 3. pricing 38%, 4. market segmentation 33%; time period Jan. 81 - Dec. 85; studies may have multiple purposes), with European conjoint studies for segmentation purposes reaching nearly the same proportions (Wittink, Vriens, and Burhenne 1994, p. 44: 1. pricing 46%, 2. new product/concept identification 36%, 3. market segmentation 29%; Jul. 86 - June 91; studies may have multiple purposes).

As for the second motivation, i.e. improving on conjoint predictive accuracy, several approaches have been put forward (see also Green, Krieger, and Schaffer 1993a, p. 345):

1. Apply empirical Bayes procedures to smooth individual-based parameters in accord with information obtained from the total sample of responses (Green, Krieger, and Schaffer 1993b).
2. "Optimal weighting" of individuals´ full profile response data with the use of Q–type factor analysis, prior to using OLS dummy-variable regression to estimate separately each person´s individual set of part-worths (Hagerty 1985, 1986, 1993).
3. Cluster respondents prior to part-worth estimation, and use the cluster-based data to maximize predictive validity (Ogawa 1987; Kamakura 1988; DeSarbo, Oliver, and Rangaswamy 1989; DeSarbo, Wedel, Vriens, and Ramaswamy 1992; Wedel and Kistemaker 1989; Wedel and Steenkamp 1989 and 1991).

The first of these approaches (Green, Krieger, and Schaffer 1993b) uses basically a self-explicated utility model with ratings of attribute-level desirabilities and attribute importances. Additionally, a limited set of full-profile stimuli, drawn from a much larger master design, is rated on a 0 to 100 likelihood-of-purchase scale. It is assumed that the best estimate of the "true" attribute-level desirabilities is found in the self-explicated desirabilities (some support for this assumptions is provided by Green and Schaffer 1991, p. 479). The full-profile stimulus ratings are used to adjust self-explicated importance weights with group-level importance weights. Thus, part-worths are only moderated by group-level importances. They are not estimated on the group level. In a pilot study, this procedure yielded no discernible advantage over the individual-level model. Thus, it is not considered for comparison in this study.

The second approach uses a Q-type factor analysis to group respondents and estimate part-worth utilities (Hagerty 1985, 1986, 1993). While Hagerty showed the capability of his approach to improve on individual-level estimation procedures with a Monté Carlo study and one empirical data set (Hagerty 1985), the only two independent empirical replications could not confirm these findings (Green and Helsen 1989; Green, Krieger, and Schaffer 1993a). While several conditions of Hagerty´s Monté Carlo study were present in the replications and could be excluded as a possible explanation for deviations, among those differences that remained were the attribute sets (i.e. type, number, levels used, correlation structure), stimulus design (i.e. fractional factorials used), and sample sizes. Hagerty, for instance, used only two profiles in the holdout sample to estimate predictive accuracy, while both replications used 16 profiles to estimate predictive accuracy. This study does not use Hagerty´s method to test improvements on conjoint with segment-level part-worth estimation.

The third type of approaches, cluster-based segmentation for conjoint analysis, have not been compared yet to other segment-level conjoint estimation methods (except for Kamakura´s hierarchical cluster analysis which has been included in Green, Krieger, and Schaffer´s 1993a comparison), or to each other. This has been accomplished by this study for the following three selected a posteriori cluster-based segmentation approaches with according tests of predictive accuracy:

(1) A hierarchical cluster segmentation method (HIC).

(2) A non-hierarchical hard clustering method (NHC).

(3) A fuzzy clustering method (FUC).

In the traditional a priori two-stage segmentation method — in contrast to above methods — subjects are first clustered into segments on the basis of characteristic variables of respondents, as for instance demographis (e.g. income, age, gender, location or channel of purchase). Thus, segmentation, here, is not based on derived benefits. Then, conjoint models are estimated for these segments, resulting in segment-level part-worth utilities. This approach may be necessary if constraints, as for instance reachability through a specific marketing channel, do not suggest an a posteriori benefit segmentation, but the marketing manager nevertheless wants to know what value structure customers of a specific channel exhibit.

Approaches (1), (2), and (3) first derive part-worths with a traditional conjoint approach, and then cluster subjects on their part-worths. After derivation of clusters, value structure is re-estimated on the segment level, and predictions for individuals are made with the part-worth model for the respective segment. Just as Hagerty (1985), Kamakura (1988) showed with one synthetic and one empirical data set that his approach with hierarchical cluster segmentation can be superior to traditional conjoint analysis. However, his finding was also not confirmed by Green, Krieger, and Schaffer´s (1993a) replication.

While methods (1) and (2) all result in non-overlapping clusters, i.e. a particular subject can only be in one, and only one cluster, fuzzy clustering allows subjects to be in a particular cluster only to a part. When comparing fuzzy cluster solutions with hard cluster solutions and Hagerty´s (1985) factor solution, patterns of partitions of subjects may be obtained as those illustrated in Table V on page 65 (adapted from Vriens, Wedel, and Wilms 1992, p. 28):

TABLE V

ILLUSTRATION OF VARIOUS PARTITIONING SCHEMES FROM SEGMENTATION

Partitioning	Non-overlapping			Fuzzy			Factor		
Clusters	1	2	3	1	2	3	1	2	3
Subject 1	1	0	0	.5	.3	.2	.8	.3	-.4
Subject 2	0	0	1	.1	.1	.8	.1	.6	.2
Subject 3	0	1	0	.1	.7	.2	-.2	.7	.8

Table V shows that for partitioning with fuzzy cluster methods the membership values of subjects in clusters sum to one over all clusters. This may be interpreted as the degree of compatibility of a particular subject with the cluster prototype. Depending on the clustering criterion used, e.g. weighted group-sum-of-squared-error (WGSS) or some graph-based method as single-linkage, different cluster solutions are possible. While the membership values of fuzzy clustering are intuitively plausible, the factor solution is difficult to interpret, which constitutes one more reason not to use it for comparison.

The fuzzy c-means algorithm (Bezdek 1981, p. 69) used for this study works similar to the non-hierarchical k–means algorithm. The crucial difference, however, is that the algorithm has one more calculation at the beginning of the comparisons: membership values in the c clusters are calculated for each data item (i.e. the subject with respective part-worth utilities as the distinguishing features) according to some distance measure. These values are iteratively adapted similar to k-means algorithm. Details may be found in Bezdek (1981) and Ruspini (1970).

The important point of all these clustering methods, however, is that meaningfulness of a cluster solution is only obtained by interpreting the cluster solution found by a particular algorithm used. In this respect, cluster solutions are like factor solutions: a cluster, or for that matter, a factor is only valid if it can be interpreted as a unit with meaning for the researcher and the objectives of the study. In terms of benefit segmentation it seems plausible to allow partial membership in segments instead of forcing a subject to belong to a particular benefit segment. One might argue the opposite for segmentation based on particular demographics: one subject may only belong to the masculine or feminine segment, but not partially into both.

It is likely an empirical question hinging on the situation which of the above approaches yields better results, individual-level or segment-level conjoint estimation. Nevertheless, it would be helpful for the marketing manager to know generalizations regarding performance and applicability of one method over the other. This study is an attempt at resolution of this question via application and comparison of cluster-based segmentation approaches and according experimental design. The tests performed to determine relative advantages in predictive accuracy for the three methods are detailed in section 3.3.6 of Chapter III.

Table VI on page 67 depicts selected conjoint studies examining effects of respondent grouping on conjoint performance (predictive accuracy and/or parameter stability). Papers above the dashed line indicate comparative studies.

TABLE VI

SELECTED STUDIES EXAMINING EFFECTS OF RESPONDENT GROUPING ON CONJOINT PERFORMANCE (PREDICTIVE ACCURACY, PARAMETER STABILITY)

Source	Date	Topic
Green and Helsen	1989	Failed cross-validation of Hagerty's 1985 Q-type factor analysis and Kamakura's 1988 hierarchical clustering
Green and Krieger	1991	Segmentation with conjoint analysis; market share estimates with 5 different segmentation strategies
Vriens, Wedel, and Wilms	1992	Monté Carlo study of five selected advanced benefit segmentation procedures for metric conjoint models
Hagerty	1993	Commentary if segmentation can improve predictive accuracy in conjoint analysis
Green, Krieger, and Schaffer	1993a	Failed replication of Hagerty's 1985 optimal respondent weighting with Q-type factor analysis for respondent grouping on three data sets
Hagerty	1985	Optimal weighting (Q-type factor analysis) for grouping results in factor solution
Hruschka	1986	Market definition and segmentation using fuzzy clustering
Ogawa	1987	Hierarchical clustering resulting in non-overlapping cluster solutions using logit model
Kamakura	1988	Hierarchical clustering resulting in non-overlapping cluster solutions
van Buuren and Heiser	1989	Discusses clustering of N objects into k groups under optimal scaling of variables
DeSarbo, Oliver, and Rangaswamy	1989	Clusterwise regression; uses non-hierarchical clustering with simulated annealing resulting in (crisp) overlapping cluster solutions
Wedel and Kistemaker	1989	Clusterwise regression; uses non-hierarchical clustering with an exchange algorithm resulting in non-overlapping cluster solutions
Wedel and Steenkamp	1989, 1991	Fuzzy clusterwise regression; uses non-hierarchical clustering with iteratively weighted least squares resulting in fuzzy cluster solutions
DeSarbo, Wedel, Vriens, and Ramaswamy	1992	Latent class procedure; optimal number of non-hierarchical clusters with an EM-algorithm resulting in fuzzy cluster solutions
Wiley	1993	General multivariate regression (GMR) for a priori segmentation; no application yet
Green, Krieger, and Schaffer	1993b	Empirical Bayes procedures to smooth individual-based parameters in accord with information from total sample

2.6 Reliability and Validity

As conjoint analysis is a relatively young discipline in marketing, concepts of reliability and validity are very diverse and not yet agreed upon among researchers in this area of science. This condition is intensified by the diverse backgrounds of researchers who are applying conjoint analysis, and who bring their respective conceptualizations of reliability and validity into this field. Bateson, Reibstein, and Boulding (1987) provided an exhaustive overview and framework for future research of conjoint analysis reliability and validity. Their framework is applied here, as it integrates diverse concepts under a common model. A more recent paper addressing this and other current issues in conjoint analysis also pointed out their framework´s usefulness for future research in this area (Green and Srinivasan 1990, p. 11). The following deliberations draw heavily on the former three authors´ ideas.

While the variety of approaches towards reliability and validity is not in itself a problem as they are applied consistently within a particular study, common approaches allow for much easier comparison across studies which also helps to stabilize and confirm the body of knowledge about value structure measurement faster than when incomparable approaches are applied. The issue prompted frequent comment but little systematic investigation. While hundreds of commercial conjoint studies are being carried out (Cattin and Wittink 1982; Wittink and Cattin 1989; Wittink, Vriens, and Burhenne 1994), the body of evidence for predictive accuracy, reliability, and validity for specific methodological approaches is rather thin and inconsistent. Many commercial studies also forego most basic reliability and validity tests and thus raise serious concerns about possible misuse. Green and Srinivasan (1990, p. 5) raised this concern in connection with widespread availability of microcomputer programs. There are mainly three issues which must be dealt with:

(1) What is the distinction between reliability and validity as identical procedures are addressed as tests of reliability by one author and as tests of validity by another ?

(2) What conceptual construct is meant with reliability in conjoint analysis ?

(3) What measures are adequate for assessing reliability and validity in this area ?

As to the first issue, Campbell (1976, p. 187; see also McCullough and Best 1979, p. 27) uses the following equation to distinguish between reliability and validity[3]:

$$X_o = X_t + X_s + X_r$$

where X_o = observed score,

X_t = true score

X_s = systematic sources of error, and

X_r = random sources of error.

A measure is called *valid* when $X_o = X_t$, i.e. when the observed score equals the true score. In contrast, a measure is called *reliable* when $X_r = 0$, i.e. when the observed score X_o does not vary due to chance or random errors and can consistently reproduce results. These characteristics make reliability a necessary, but not sufficient, condition for validity (Bateson, Reibstein, and Boulding 1987). Peter (1979, p. 7) counts systematic sources of variance into the true score, and argues that distinctions between systematic variances and true variances are not an issue of reliability but one of (construct) validity. Campbell and Fiske (1959, p. 81) argue for separation of true

[3] His notation is $X = T + e_s + e_c$ which is re-written here in a more mnemonically amenable way.

score and systematic error, as systematic variance among test scores can be due to responses to the measurement features (e.g. via order effects) as well as responses to the trait content, and may be detected when examining test score correlations.

With the basic conjoint model as written in equation (E3.1.4) in Chapter III on page 81, the observed, respectively derived, part-worth utilities b_{ij} are valid when they represent true measures of the respondent´s underlying part-worths for the i-th level of the j-th attribute. Also, they are reliable when they contain no variation due to random factors but may contain variation due to systematic error.

However, in making both concepts operational, the distinction between validity and reliability blurs because we are not able to obtain a measure of X_t, the true score, only surrogates or approximations to it. A widely used approach, the multitrait multimethod matrix suggested by Campbell and Fiske (1959) provides considerable insight (see also Churchill 1979, p. 69). A correlation matrix of different traits measured in different ways is used to assess validity. The researcher looks for correlations between tests intended to measure the same trait (convergent validity) and no correlation between tests intended to measure different traits (discriminant validity). Citing Campbell and Fiske (1959, p. 81), "tests can be invalidated by too high correlations with other tests from which they were intended to differ." Bateson, Reibstein, and Boulding (1987, p. 454) conclude that "all conjoint studies have focused on convergent rather than discriminant validity. Indeed, it is difficult to see how discriminant validity could be applied to conjoint analysis."

While Campbell and Fiske (1959) define convergent validity as agreement between two attempts to measure the same trait with maximally different methods, they define reliability as agreement between two efforts to measure the same trait with maximally similar methods (1959, p. 83). Therefore, the distinction between the concepts of reliability and convergent validity hinges on operational definitions of "maximally similar" and "maximally different." For this, however, there is no definite answer because in reality a spectrum extends from reliability to convergent validity (Campbell and Fiske 1959, p. 83). Thus, identical procedures can be used legitimately to test convergent validity and reliability, depending on the researcher´s definition. This study follows Bateson, Reibstein, and Boulding´s view (1987, p. 454) that most conjoint studies perform only reliability checks when the checking task takes an additional decompositional approach based on active evaluation experiments. Therefore, the only checks that qualify as validity checks are those that compare conjoint analysis results with behavior or with self-explicated importance and part-worth weights, with the latter constituting a weaker test of validity than the former (Leigh, MacKay, and Summers 1981, p. 321), but one that is often the only possible check to resort to. This study uses a self-explicated model for determining (convergent) validity.

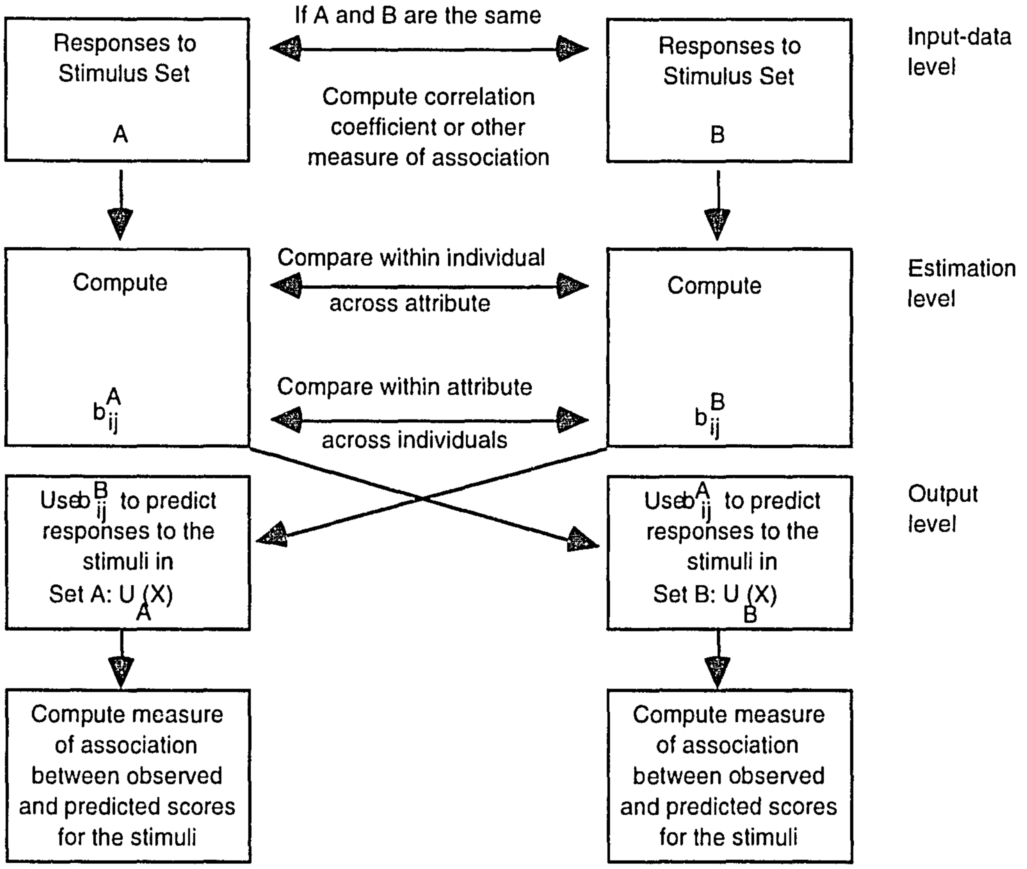
Considering the second issue of conceptual form of reliability, much of reliability research assumes a single construct (Churchill 1979, p. 69). In contrast, this study takes Bateson, Reibstein, and Boulding´s (1987, p. 455) view from generalizability theory that there is no such thing as a single reliability score; "rather the score must specify the conditions of measurement over which reliability has been measured," i.e. reliability is context-dependent. This may also be termed the systems view towards reliability[4]. Therefore, reliability in connection with conjoint experiments may be classified as reliability over time (Leigh, MacKay, and Summers´ 'temporal reliability' 1981, p. 318; also McCullough and Best 1979, p. 27 'temporal stability'), reliability

[4] A more detailed discussion of the systems view towards reliability and performance measures is provided later in this study.

over attribute set, reliability over stimulus set, and reliability over data collection method.

Reliability over time is assessed when the only aspect varied is the time of administration of the conjoint experiment. Everything else is held constant. The question is whether b_{ij}´s at time t are the same as those at time t + lag where lag is some time lag. Reliability over attribute set is assessed when the stability of part-worths for a common (core) set of attributes is examined as other attributes within stimuli are varied. It is achieved when part-worths for a given attribute level and a specific individual do not depend on the presence of other attributes. Therefore, this test may also be viewed as a test of the additive model, the value structure without interaction effects, or as a test of the hypothesis of independence from irrelevant attributes (IIA; Oppewal, Louviere, and Timmermans 1994, p. 104). Reliability over stimulus set assesses whether derived part-worths are sensitive to the fractional factorial design used for estimation, i.e. to subsets of profile descriptions. This problem is absent in full factorial designs. However, as judgmental limitations on the stimulus set are always uneasy compromises with potential distortions, it should at least be certain that possible distortions do not emanate from the factorial design chosen. Reliability over data collection method consists of three aspects: type of data, data-gathering procedure, and type of dependent variable. If part-worths differ depending on variations in any of these data collection methodologies respective part-worth utilities cannot be relied on. This study tests for reliability over time, over attribute set, and reliability over stimulus set.

Finally, concerning appropriate measures for reliability, there is no agreement in the literature. Even worse, new measures are added without providing exhaustive rationales for inappropriateness of existing ones, concerning for example the type of situation, type of product, or other corrollaries. For studies with full replications, i.e. with the same set of items, the following measures have been applied: the R^2 ratio, the Pearson product moment correlation of the estimated part-worths across respondents but not attributes, or across attributes but not respondents, comparison of the input data (i.e. profile scores), and several measures based on distance between the b_{ij}'s. Figure 6 on page 73, adapted from Bateson, Reibstein, and Boulding (1987, p. 458), illustrates the possibilities and connection between measures.



Responses to Stimulus Set A
If A and B are the same
Compute correlation coefficient or other measure of association
Responses to Stimulus Set B
Input-data level
Compute
b_{ij}^{A}
Compare within individual across attribute
Compare within attribute across individuals
Compute
b_{ij}^{B}
Estimation level
Useb b_{ij}^{B} to predict responses to the stimuli in Set A: $U_A(X)$
Useb b_{ij}^{A} to predict responses to the stimuli in Set B: $U_B(X)$
Output level
Compute measure of association between observed and predicted scores for the stimuli
Compute measure of association between observed and predicted scores for the stimuli

Figure 6. Alternative Measures of Reliability (adapted from Bateson, Reibstein, and Boulding 1987, p. 458).

Measures of reliability may be obtained at different levels into the process of performing a conjoint study, mainly at the input-data level, the estimation level, and at the output level. At the input-data level, some correlation measure compares the overall utility results from two administrations in their plain or some adjusted (e.g. standardized) form. Measures that have been used are

- Pearson product moment correlation, and
- rank correlation coefficients.

As for the estimation level,

- b_{ij}´s may be computed separately for different samples or groups and compared with each other,
- b_{ij}´s from one half of the study may be used to predict utilities of the stimuli of the other half, comparing predicted with actual overall utilities with aforementioned measures of association across attributes or across respondents/individuals.

At the output level, cross-validations with holdout samples may be performed on additional stimuli using the original design, holdout stimuli from a separate design, or replications from the original design (the latter two approaches are applied in this study). An additional approach both at the estimation and at the output level is "jackknifing", which involves estimation of b_{ij}´s leaving one observation out, respectively, and observing how stable the estimations of part-worths are, or how stable predictions of overall utilities are (which is similar to studentization of t-values). This procedure has the advantage not to require additional data for testing. Here, measures used have been

- rank correlations,
- product moment correlations of observed and predicted scores for the holdout (applied in this study),
- ability to predict the most and least perferred stimuli, and
- number of first hits, i.e. the stimulus chosen out of a set of stimuli (also applied in this study).

As Bateson, Reibstein, and Boulding (1987, p. 459) point out, the properties of these measures are as yet unknown, and it is not known which measure is the most appropriate for which kind of study. They suggest the consideration of three factors in selecting a measure:

(1) the reliability of what shall be measured,

(2) what does significance mean in this context, i.e. what is the statistical power, and

(3) what data requirements are there to use one measure ?

For assessing reliability of value structure model form, overall utilities of the stimuli U(X) must be compared, and a measure of reliability be applied. For assessment of reliability of value structure itself, i.e. reliability of part-worths, measures must compare part-worth utilities, i.e. b_{ij}´s. For segmentation and (new) product policy decisions, reliability of part-worth utilities is of utmost importance, while for choice and market share predictions, stable overall utilities are most important. Due to compensatory effects it may turn out that overall utilities are more reliable than part-worth utilities. However, Leigh, MacKay, and Summers (1981, p. 318) argue that the higher degrees of freedom (DFs) generated by examining stimulus utilities are "partially illusory since these values are functionally related through the part-worths." This study examines reliability for both types of utilities in the cases of reliability over time and stimulus set, as well as in the case of reliability over attribute set (cp. overview of tests in Table XI on page 131).

The last argument of Leigh, MacKay, and Summers (1981) focuses discussion on the believability of statistical tests in connection with utility estimates. It may be emphasized, that the low ratio of observations per parameter estimated raises serious doubts about significance tests based on these numbers. In many other areas of statistical measurement theory, observation to item ratios of 8:1 to 10:1 are deemed sufficient to support confidence levels based on results of Monté Carlo studies, like for instance in (confirmatory) factor analysis. To this author's knowledge, no Monté Carlo study has been performed yet to establish similar rules of thumb for significance tests of utility estimates in conjoint analysis[5]. This study does not use significance tests of part-worth utilities on the basis of one individual for generalizations about appropriate model form, but attribute levels were generally plotted against each other and visually checked for interactions, i.e. for those interactions included in the design (Louviere 1988, p. 20 and p. 33).

As for data requirements concerning reliability measures, part-worth utility measures necessitate at least one complete replication, preferably more. Without considerable incentives respondents are not willing to perform such a task, and perform it in a useful, careful manner. On the other hand, reliability measures of the dependent variable do only require additional observations from respondents with possibly different designs. This allows for measures that yield insights beyond mere predictive accuracy of one model. Therefore, this is the approach taken in this study. Rather than arguing for arbitrary (and ultimately indefensible) cutoff points for reliability

[5] There is, however, now a Monté Carlo study by Umesh and Mishra (1990) that establishes rules of thumb for reliability of selected performance measures (index-of-fit) of related respective conjoint procedures (programs).

measures, as for example 0.7 for Pearson´s product moment correlation coefficient, Bateson, Reibstein, and Boulding (1987, p. 455) plead for a practical view of the problem of reliability by asking whether a procedure chosen is any more reliable than available alternatives, as for instance self-explicated attribute (level) utilities which, at the same time, is a check on (convergent) validity. This is the approach taken in this study, though two more, respectively three more reliability measures are calculated than "necessary" (First-Hit, RMSE, and R^2) in order to make this study comparable with prior and future studies.

In summary, even seven years after Bateson, Reibstein, and Boulding´s comprehensive review (1987) of conjoint reliability and validity their conclusion still seems to hold (p. 477): "In developing our review, we had hoped that a synthesis of the literature would afford insights into the best conjoint analysis procedure and the most appropriate methodology to use for assessing reliability and validity. Instead, we have highlighted just how little is known about these areas." This study contributes to the compilation of additional knowledge in this area of conjoint analysis.

Table VII on page 78 lists selected conjoint studies examining reliability and validity.

TABLE VII

SELECTED STUDIES WITH EMPHASIS ON CONJOINT RELIABILITY AND VALIDITY

Source	Date	Characteristic
McCullough and Best	1979	Early discussion of the multidimensionality of reliability in ConjA; distinction between temporal and structural reliability (i.e. over time and stimuli)
Bateson, Reibstein, and Boulding	1987	Complete review of conjoint reliability and validity studies until 1984; develop conceptual organization of reliability and validity as applied to ConjA
Reibstein, Bateson, and Boulding	1988	Empirical findings for reliability over attribute set and over stimulus set for five product categories
Wittink, Reibstein, Boulding, Bateson, and Walsh	1989	Compare use of alpha, i.e. the probability of obtaining a sample result under H0 of perfect agreement in two parameter vectors, with correlation for part-worths (both are dependent on the number of part-worths compared)
Umesh and Mishra	1990	Monté Carlo investigation of three ConjA index-of-fit measures (C*, stress, and R^2)
Hagerty and Srinivasan	1991	Comparison of predictive power of alternative multiple regression models; as analogy for model choice of conjoint models as parameter-dependent

CHAPTER III

METHODOLOGY, RESEARCH QUESTIONS, AND PROCEDURES

This chapter is composed of three parts: methodology, research questions, procedures and descriptions of the data needed. First, conjoint models and conjoint methodology are described as they are applied in this study. Next, research questions and related hypotheses addressed in this study are presented as they have been derived from the literature review. Finally, a description of the procedures and data for measurement is provided.

3.1 Methodology

In this section, value measurement models are presented as they are applied in this study. First, general model representation and related terminology is introduced. Then, specific conjoint model forms are illustrated.

3.1.1 Conjoint Analysis Models

General Model Representations

Without regard to the preference or choice elicitation technique used to empirically assess value structure, a variety of models have been used to characterize customers´ multiattribute utility functions. Each alternative or choice option **X** is represented as an ordered M-tuple of M decision attributes:

$$\mathbf{X} = (x_1, x_2, \ldots, x_m) \qquad \text{(E3.1.1)}$$

where $x_1, x_2, \ldots, x_j, \ldots, x_m$ refer to the level (or position or state) of the j-th attribute describing **X**.

If an attribute is categorical (i.e. its p "levels" or positions are unordered) it may be coded non-redundantly in the alternative in form of p-1 dummy variables.

The value or utility function from differential evaluations of attribute positions by the decision maker may be expressed as:

$$U(x_1, x_2, \ldots, x_j, \ldots, x_m) = f\,[u_1(x_1), u_2(x_2), \ldots, u_j(x_j), \ldots, u_m(x_m)] \qquad \text{(E3.1.2)}$$

where each u_j is a part-worth function defined over all values of the j-th attribute. These part-worth utility functions $u_j(x_j)$ may be constrained to have linear, quadratic, or other functional forms for all levels of attribute j, or they may be unconstrained. $f\,[\cdot]$ denotes a function that aggregates part-worths over the attributes. In the notation of the conceptual model in Figure 3 on page 17 and of equation (E2.3.2) on page 48, $u_j(x_j)$ comprise mappings f_1 and f_2, and $f\,[\cdot]$ in equation (E3.1.2) comprises mappings f_3 and f_4. $u_j(x_j)$ are the attribute values (i.e. utilities) $V(S_{nj})$ of the conceptual model.

Additive Models

The most frequently used model for aggregation of part-worths is the additive model in form of two-stage, self-explicated utilities for attribute levels (i.e. $u_j(x_j)$) and importance weights (i.e. w_j) for these attributes:

$$U(x_1, x_2, \ldots, x_m) = \sum_{j=1}^{m} w_j u_j(x_j) \qquad \text{(E3.1.3)}$$

where

$$\sum_{j=1}^{m} w_j = 1.0$$

This additive model is not (truly) a conjoint model, but an expectancy value model which is used in conjunction with conjoint models when limited numbers of observations due to large numbers of attributes do not allow for pure derivation of part-worths (details of this problem are provided later).

The corresponding (main effects only) additive conjoint model with part-worth utilities derived by means of some decompositional, regression-like procedure is denoted as follows:

$$U(x_1, \ldots, x_i, \ldots, x_m) = b_0 + \sum_{j=1}^{m} \sum_{i=1}^{L_j} b_{ij} d_{ij} \qquad \text{(E3.1.4)}$$

where b_0 denotes the intercept (if non-zero),

b_{ij} is a partial regression coefficient,

d_{ij} is a dummy variable with 1 if attribute j is at level i, and 0 otherwise, and

L_j denotes the number of levels for attribute j.

Depending on the constraints put on the part-worth utility functions $u_j(x_j)$ over respective attribute levels, and depending upon according coding, b_{ij}'s represent two or more part-worth utilities. In order to derive the incremental contribution of the i-th level of the j-th attribute, i.e. the α_{ij}'s part-worth, towards overall utility $U(\mathbf{X})$, several sets of equations have to be solved, simultaneously. In the case of three-level attributes, and no constraints put on part-worth utility model form (as applied in this study), the following three sets of equations have to be solved for each attribute j:

$$\sum_{i=1}^{3} \alpha_{ij} = 0 \qquad \text{(E3.1.5)}$$

$$\alpha_{1j} - \alpha_{3j} = b_{1j} \qquad \text{(E3.1.6)}$$

$$\alpha_{2j} - \alpha_{3j} = b_{2j} \qquad \text{(E3.1.7)}$$

where j = 1, 2, 3 (or more generally, j = 1, 2, ..., m).

A noteworthy difference to the self-explicated additive model is that importance of the attribute and respective level utilities are not estimated separately, but ensemble in the b_{ij} coefficients. Therefore, after estimation of part-worths, importance weights must be computed in an additional step. Relative importances of attributes are computed with the following equation:

$$w_j = \frac{[\operatorname*{Max}_i(\alpha_{ij}) - \operatorname*{Min}_i(\alpha_{ij})]}{\sum_{j=1}^{m}[\operatorname*{Max}_i(\alpha_{ij}) - \operatorname*{Min}_i(\alpha_{ij})]} \quad \text{, for each j} \qquad \text{(E3.1.8)}$$

where $[\operatorname*{Max}_i(\alpha_{ij}) - \operatorname*{Min}_i(\alpha_{ij})]$ denotes the range of part-worths over all levels i of attribute j.

The expectancy value model of equation (E3.1.3) is the base model against which predictive performance of the individual-level (traditional) conjoint model is compared in Phase I of this study. Both these types of models are compared to segment-based conjoint models in Phase II of this study.

Multilinear Utility Models

The most flexible conjoint model allows multiple interaction terms among attribute levels for representation of various forms of nonlinearity in (part-worth) utility aggregation:

$$
\begin{aligned}
U(x_1, \ldots, x_i, \ldots, x_m) = \quad & b_0 + \sum_{j=1}^{m} \sum_{i=1}^{L_j} b_{ij} d_{ij} \qquad \text{(E3.1.9)} \\
& + \sum_{i \neq k} b_j d_j * b_k d_k \\
& + \sum_{i \neq k \neq r} b_j d_j * b_k d_k * b_r d_r \\
& + \ldots \text{ (all other possible interactions)}
\end{aligned}
$$

where all terms of the first row are equivalent to the additive conjoint model, the second row denotes pairwise interaction terms between attribute levels, the third row denotes triple interaction terms among attribute levels, and so on until all possible interactions are represented in the model.

In practice, however, researchers rarely go beyond models of selected two-way interaction terms (Green and Krieger 1993, p. 471). In commercial studies, modeling of interaction terms is virtually absent, though a majority of applied researchers acknowledge their importance. Wittink, Vriens, and Burhenne (1994, p. 50) report that only 10 percent of commercial studies include interaction terms.

Hybrid Models

Hybrid conjoint models combine self-explication of attribute levels and attribute importances with decompositional conjoint models. The most used hybrid main-effects-only model is represented as

$$U(x_1, \ldots, x_i, \ldots, x_m) = a + b\sum_{j=1}^{m} w_j u_j(x_j) + \sum_{j=1}^{m}\sum_{i=1}^{L_j} b_{ij} d_{ij} \qquad \text{(E3.1.10)}$$

where a denotes the intercept (if non-zero),

b is a regression coefficient that represents the contribution of the self-explicated term to $U(x_j)$,

$u_j(x_j)$ is the utility of the level of the j-th attribute,

w_j is the importance of attribute j,

b_{ij} is a partial regression coefficient,

d_{ij} is a dummy variable with 1 if attribute j is at level i, and 0 otherwise, and

L_j denotes the number of levels for attribute j.

Hybrid conjoint models have been developed to reduce the burden imposed on respondents when the number of required evaluations increases due to a large number of attributes and their respective levels, but to still allow individual-level utility functions. For this model, respondents provide self-explicated utilities for all attributes while responding only to a small number of stimulus profiles. Then, the self-explicated utilities are combined with utilities from a conjoint analysis which has been estimated across a number of respondents.

3.1.2 General Design and Estimation Considerations

There are generally two approaches to measuring the dependent utility variable: as a rank-ordered or as an interval-scaled rating variable. Ranking involves data collection methods which present respondents with at least two attributes or profiles at a time, and the procedure can become quite unwieldy with a large number of attributes. Rating procedures ask respondents to rate a particular profile on some form of preference or behavioral intention scale. While there is still some disagreement whether subjects´ responses may be more accurately recorded on a ranking or rating scale, rating scales and dummy variable regression are reported to be the most widely used methods, given that comparisons of both methods and associated estimation procedures did not yield substantially different results (Jain, Acito, Malhotra, and Mahajan 1979, pp. 318; Green and Krieger 1993, p. 478). Rated overall utilities and OLS regression are also the methods of choice in this study.

Due to the number of levels and attributes in this study, it is necessary to employ a highly fractionated experimental design. Details of the design and analyses are provided in later sections of this chapter.

3.2 Research Questions

The research questions as stated in the introduction and supported by the literature may be summarized as follows. Related hypotheses and their testing procedures are provided in section 3.3.6 on pp. 125.

1) What is the influence of the type of attribute chosen for the evaluative task (i.e. technical or product-referent attributes and non-technical or user-referent attributes) on customer value structure and predictive validity ?
2) What is the influence of specific factorial designs, i.e. of specific combinations of product attribute values, on estimation of customer value structure and predictive accuracy ?
3) How do type of attribute in the product profile and factorial design interact in their influence on customer value structure for different models ?
4) Which individual-level model for customer value structure performs best with respect to prediction ?
5) Can cluster-based segmentation approaches improve accuracy in prediction of value attributions to product profiles over individual-level conjoint models ?
6) Which aggregate model for customer value structure performs best with respect to prediction ?
7) Are the purposes of prediction and segmentation, as well as potential other purposes, better served with the suggested methods, and what practical limitations are there for the different methods to support specific purposes ?
8) Are benefit segments obtained with different clustering procedures meaningful for target marketing, or may they only increase predictive accuracy ?

3.3 Data and Procedures

First, a general account of study design is provided, describing all elements of the behavioral system to establish the framework for more detail. This is accomplished by illustrating the two phases of study design. Then, data type and sources are described. Finally, procedures for analysis and general outlines of expected results are illustrated.

3.3.1 Experimental Design

The design of the study involves two phases:

(1) In Phase I, effects of methodological variations in conjoint on observed benefit and utility measures are traced for the self-explicated and the traditional (individual-level) conjoint model. A comparison between both types of models establishes (convergent) validity for an individual´s utility measures.

(2) In Phase II, the focus of this study, different segmentation methods are used to group subjects into meaningful segments, and to assess improvements on conjoint predictive accuracy and reliability.

Analyses performed at each stage of the study are detailed in the next section. Figure 7 on page 88 illustrates the phases of the study as they pertain to timely procedure.

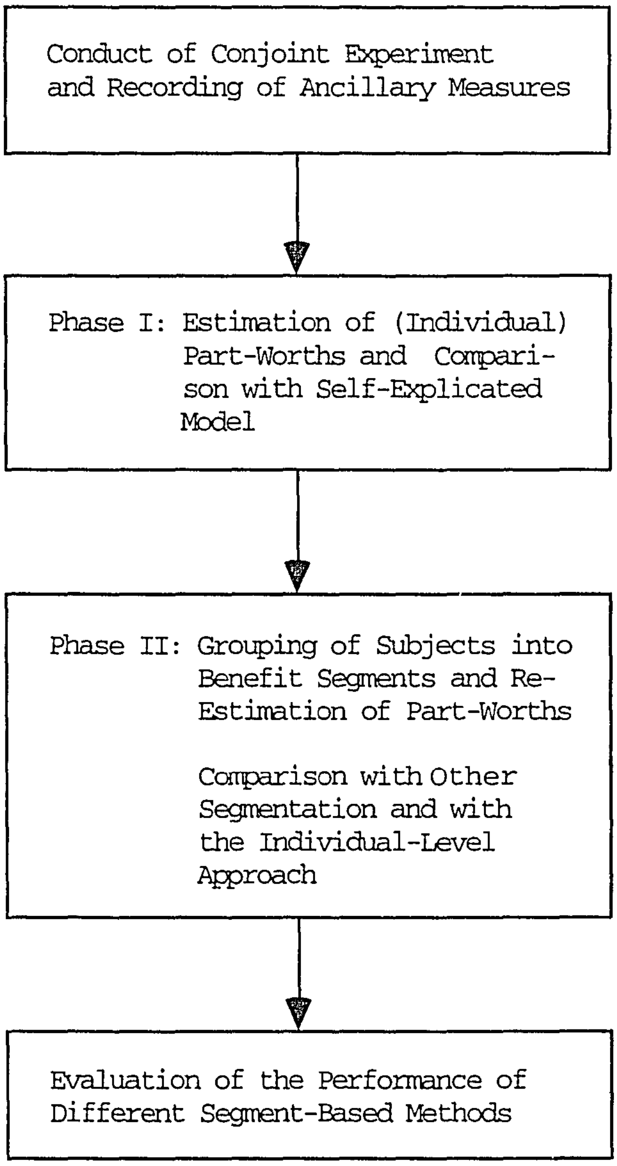

Conduct of Conjoint Experiment and Recording of Ancillary Measures
Phase I: Estimation of (Individual) Part-Worths and Comparison with Self-Explicated Model
Phase II: Grouping of Subjects into Benefit Segments and Re-Estimation of Part-Worths
Comparison with Other Segmentation and with the Individual-Level Approach
Evaluation of the Performance of Different Segment-Based Methods

Figure 7. Phases of Research Study.

The choice of experimental manipulations reflects effects that were either found or suggested to have major impacts on the estimation of customer value structure in the literature review. In particular, they reflect most recent suggestions emanating from only two limited empirical validations of only two of the new conjoint segmentation methods (two independent validations of Hagerty´s 1985 Q-type factor analysis approach and Kamakura´s 1988 hierarchical clustering approach). The design strategy is to maintain as few confoundings of effects as possible (e.g. variability in measurements over time and over different stimulus sets), in contrast to a design strategy that deliberately confounds effects assumed to point in the same directions (i.e. effects expected to increase variability in measurement are confounded to increase measurement contrast). The design, therefore, allows for tracing of selected methodological influences on part-worth estimates (i.e. value structure), reliability, and predictive validity. Particularly, it allows for measurement of influences of methodological variants of conjoint analyses on purposes of prediction and segmentation, as well as on related measures.

For Phase I, the experimental design is a repeated measure posttest-only, 2 x 2 x 5 design with two levels of attribute types (strictly product-referent or technical attribute set, and mixed technical and user-referent attribute set), two levels of stimuli sets (first fractional factorial stimuli set, and second fractional factorial stimuli set), and five levels of model form (one self-explicated and four conjoint models).

For Phase II, the experimental design is a repeated measure posttest-only 2 x 2 x 3 factorial with the same first 2 x 2 as before, but then with three levels of segmentation approaches (hierarchical clustering, non-hierarchical hard clustering, fuzzy clustering). Table VIII on page 90 and Table IX on page 91 are representations of the respective design layout. The design layout considering administration of measurements is given in Figure 8 on page 92.

TABLE VIII

PHASE I: EXPERIMENTAL DESIGN FACTORS FOR INDIVIDUAL-LEVEL ANALYSES

Type of Attributes	Factorial Design of Stimulus Set	Model Form	
		Self-Explicated (SE)	Individual-Level Conjoint (TC)
Product-Referent or Technical Attributes (A1)	Fractional Factorial 1 (FF1)	R^2, First-Hit, etc.	R^2, First-Hit, etc.
	Fractional Factorial 2 (FF2)	R^2, First-Hit, etc.	R^2, First-Hit, etc.
Product-Referent and User-Referent Attributes (A2)	Fractional Factorial 1 (FF1)	R^2, First-Hit, etc.	R^2, First-Hit, etc.

Note: R^2, etc. denote the performance measures of respective models (pp. 106).

TABLE IX

PHASE II: EXPERIMENTAL DESIGN FACTORS FOR SEGMENT-LEVEL ANALYSES

Segmentation Method	Number of Clusters	Type of Attributes and Factorial Design of Stimulus Set (A1, A2, FF1, and FF2 Pooled)
Hierarchical Clustering (HIC)	3 Clusters	R^2, First-Hit, etc.
	4 Clusters	R^2, First-Hit, etc.
Non-Hierarchical Hard Clustering (NHC)	3 Clusters	R^2, First-Hit, etc.
	4 Clusters	R^2, First-Hit, etc.
Fuzzy Clustering (FUC)	5 Models with 3 Clusters	R^2, First-Hit, etc.
	5 Models with 4 Clusters	R^2, First-Hit, etc.

Note: R^2, etc. denote the performance measures of respective models (pp. 106).

1st Replication ——> time line ——> 2nd Replication

R1	A1FF1	OS1	R11	A1FF1	OS3
(G1)		OC1	(G3)		OC3
		OD1			OD3
		OH1			OH3
			R12	A1FF2	OS4
			(G4)		OC4
					OD4
					OH4
R2	A2FF1	OS2	R21	A2FF1	OS5
(G2)		OC2	(G5)		OC5
		OD2			OD5
		OH2			OH5
			R22	A1FF1	OS6
			(G6)		OC6
					OD6
					OH6

Figure 8. Design Layout Concerning Administration of Measurements.

where R denotes random group assignment of subjects to treatments with groups in brackets as applied in the results section of Chapter IV,

A1 is the product-referent or technical attribute set,

A2 is the mixed (user-referent and technical) attribute set,

FF are fractional factorials,

OS are the observations of self-explicated measures,

OC are the observations of conjoint stimulus evaluations

OD are the recordings of demographic variables, and

OH denote the observations of holdout stimuli.

3.3.2 Research Variables of Phase I

Selection of Product

The product chosen for evaluation in this study is a laptop, notebook, or portable computer. It is a tangible, durable business and consumer product which may reasonably well be characterized as a high-involvement product where the assumption of compensatory decision rules are well-documented in the consumer research literature. It is relatively new and complex, is still relatively expensive, and satisfies diverse customer needs. These different needs may provide favorable conditions for divergent benefits attributed to the product's characteristics. These different benefit attributions expressed in different part-worths may then be useful candidates for segmentation strategies. Furthermore, many young people and especially current student population are quite familiar with at least its immobile counterpart, a desktop computer.

Part of this study's research objective is to examine the question if conjoint analysis is also a valid measurement tool for an as innovative, technically complex, and rapidly evolving consumer product as notebooks are. These product characterizations are accurate for an increasing number of technologically oriented consumer products, as for instance in consumer electronics. Familiarity with these innovative products is not as high as with some other technical products, for example cars, or as with many non-technical products, as for instance food, beverages, or apartments. However, some familiarity with the product must be present in order to keep the assumption of compensatory decision rules as good approximators of the customer decision process.

While more innovative products or truly new product categories, like for instance PDAs mentioned in the introduction, may provide a better basis for divergent customer segments, low familiarity with the product may lead to decision strategies that are not attributable to product characteristics but to the consumer or referent others, as for instance family members, friends, or colleagues. A pretest concerning the importance of decision criteria in determining the purchase likelihood for laptop or notebook computers increased confidence that the product evaluation is not based on referent others but mainly on product characteristics. Details of the pretest are provided in Appendix I.

Finally, a laptop is a technical product, the characteristics of which can be described with mainly monotone attribute levels. This is an important characteristic, as conjoint analysis works best where consumption decisions are based on value attributions towards particular product characteristics, in contrast to purchase decisions that are made wholistically, as for example on the basis of aesthetics. As interactions with computers in business and private life are ever more inevitable one needs to know how people make value judgments for these products as opposed to less technical products, and one needs to know in which contexts a particular measurement model is applicable for marketing purposes. Therefore, a laptop or notebook computer satisfies the major criteria for inclusion in this study.

Selection of Attributes and Levels as Independent Variables

For conjoint analysis to work it is important to understand the decision problem and its environment faced by target individuals. It works best when all key determinant decision attributes are identified. However, the inclusion of particular attributes is always an uneasy compromise between strifing for completeness of the relevant decision criteria and keeping the evaluation task in line with respondent capabilities. Furthermore, decision attributes should be as amenable to managerial manipulation as possible, i.e. they should be actionable and measurable. Therefore, the trade press was perused and informal interviews were conducted with computer users and non-users to identify relevant decision criteria. Additionally, one informal interview was conducted with sales reps and the manager of a local computer store. This information was then condensed and attributes and their respective levels were chosen so that they denoted broad categories of choice criteria. Levels were chosen so that metric variables comprised the extreme values of current, most widely available real products. Levels of metric attributes were evenly spaced, and nonmetric levels were chosen to imply an order.

A pretest was conducted to elicit the stated importance of ten candidate attributes with the intention to narrow down this list to about six to eight at two or three levels which is considered to be a good balance between demands for conjoint design and realism of respondent task before one may experience simplified decision strategies. The pretest also encouraged to state criteria a respondent would use but that were not included in the importance ratings. Additionally, one control variable, familiarity with the product class, was rated, and another control, the order of questions on the questionnaire, was obtained. Based on the results of the pretest, it was decided to drop only one attribute from the final list, add one technical attribute that can be exchanged with the non-technical attribute, and keep the other nine (9) attributes for the main study. Additional idiosyncratic decision criteria obtained with the last question on the pretest questionnaire resulted in no discernible broad categories in addition to the stated ones that may have been overlooked.

This yielded eight (8), respectively nine (9), technical or product-referent attributes, and one (1) user-referent attribute. A check for order effects in the questionnaire items did not reveal significant effects, though some may be considered borderline cases. A check for the presence of negative attribute correlations did not reveal severe conditions. The presence of nominal and metric types of attribute levels may have indicated increased potential for interaction effects. However, this could only be confirmed in tests in the main study (actually, interactions were not significant on the group level; cp. Chapter IV), and was screened for in plots of attribute level utilities against each other. Finally, a covariance analysis was conducted using familiarity with the product class in order to elicit this ancillary variable´s potential for revealing differentiating benefit attributions of respondents (i.e. act as a control variable for consumer differences), and thus serve as a segmentation base. Though not significant, a visual inspection suggested a potential for those controls to serve as useful segmentation bases. Details of the pretest and related analyses are provided in Appendix I. Table X on page 99 provides an overview of the attributes and levels used for this study.

Dependent Variable 'Purchase Likelihood'

Purchase likelihood was obtained on a rating scale ranging from 0 (definitely would not buy this notebook computer) to 100 (definitely would buy this notebook computer). Respondents were asked to imagine they were in the situation of evaluating different laptop computers for future purchase as their own computer. It was obtained by asking respondents to rate a product profile by distributing a number of points ranging from zero (0) to one hundred (100) to the profile being evaluated, denoting his/her stated likelihood of purchase for the given attribute level combinations describing one specific stimulus (i.e. laptop computer). Likelihood of purchase was chosen over preference or desirability because it is assumed to be the better term to denote preference with respect to a buying situation (i.e. reminds respondents of the situation in which the evaluation takes place) and thus nearer as a surrogate to market behavior than the latter two terms (see also Green and Schaffer 1991, p. 477).

Ancillary Variables

The following ancillary variables as candidates for potential covariates and their respective scale types are included in the study: Familiarity with the product class (category rating scale), time to complete the experiment (minutes from start of the experiment), perceived difficulty of the evaluation tasks (category rating scale), gender (binary), age (number of years), year as undergraduate (freshman, sophomore, junior, senior) or graduate, years of work experience, computer ownership (no-yes[years]), computer usage and experience (number of years).

As it was possible in this study to identify respondents, some desirable ancillary variables that are commonly found to provide good differentiators among individual consumption behaviors (for segmentation) were not recorded, as for instance demographics like income, or psychographic construct items to identify lifestyles; the danger of biased answers did not make it worthwhile (Montgomery 1986; however, cp. increased predictive accuracy of combined attribute, i.e. conjoint, and LOVs, i.e. list of values, models in a recent study by Sukhdial, Chakraborty, and Steger 1995, Fig. 1, p. 16). Familiarity is included because familiarity with the product class is assumed to be directly related to ability of performing the respondent task, and, in its absence, responsible for high variance in the ratings or breakdown of the conjoint task (for a distinction between familiarity and knowledge and its significance for performing value judgments see Alba and Hutchinson 1987). Familiarity, as well as other demographics collected can reasonably be assumed to be non-biased responses, given that the experiment was confidential, though not anonymous. A similar justification is provided for recording of perceived difficulty of the evaluation task. Due to data collection procedures as self-administered questionnaires and due to many missing values, time to complete the experiment is judged to be too unreliable to provide sufficient basis for segmentation. It was dropped from subsequent analysis though it may indicate outliers in terms of care with which the evaluation task has been performed. The rest of above ancillary variables, as well as familiarity, are conjectured to provide a reasonable basis for user-related and product-experience-related segmentation, and were used to cross-tabulate with the HIC segments found. However, no significant differences were identified. Therefore, no further cross-tabulations were performed.

TABLE X

OVERVIEW OF ATTRIBUTES AND LEVELS USED IN CONJOINT STUDY

	Attribute	Levels	Characteristics (Visualization)
1. (D)	Weight	9 pounds[a] 7 pounds 5 pounds[a]	monotonic; metric
2. (E)	Screen Size	8.4 inch (diagonal)[a] 9.4 inch (diagonal) 10.4 inch (diagonal)[a]	monotonic; metric (show sheets of paper in actual size)
3. (C)	Display Type	Monochrome Color	nominal
4. (H)	Base Price	$ 3500[a] $ 2500 $ 1500[a]	monotonic; metric
5. (B)	Keyboard Size	Smaller than regular size Regular size	nominal
6. (F)	Battery Life	3 hours[a] 5 hours 7 hours[a]	monotonic; metric
7. (A)	Performance/Speed	Comfortable for word-processing Fast for big spreadsheet and imaging	ordinal
8. (G)	Feature Load	No additional features[a] Expansion slots for key-board, monitor, others Faxmodem, CD-ROM, expansion slots for key-board, monitor, others[a]	ordinal
9. (I)	Pointing Device	Mouse[a] Trackball Trackpad or other[a]	nominal
10. (I)	Firm Reputation (Brand)	No-name[a] Store brand Well-known brand[a]	nominal

[a] Levels used for the 2-level extreme design of the holdout product profiles.

(.) Letters in brackets denote attribute order before randomization, and as identified in the model form.

?. Figures in front of attributes indicate their order on the questionnaire (and thus the reverse order, too).

3.3.3 Research Variables of Phase II

This study uses four independent variables and three, for some tests five, dependent variables of major importance. The independent variables are type of attributes (A1, A2), type of factorial design (FF1, FF2), model type (SE, TC), and segmentation method (HIC, NHC, FUC). The dependent variables are the coefficient of determination (R^2), the adjusted form of R^2 (Adj R^2), root mean squared error of prediction (RMSE), Pearson´s product moment correlation coefficient (r_{xy}), and first choice hit rate (First-Hit) as surrogate measures of predictive performance (purchase likelihood). In addition, several ancillary variables were measured as potential covariates. They are explained subsequent to product attributes for the traditional (i.e. the base) conjoint experiment.

Independent Variables

Type of Attribute Set.

There are two types of attribute sets to be evaluated, A1 and A2, which differ in the types of attributes used to describe the product. The number of attributes (nine per stimulus description) and the levels within attributes (two or three per attribute) remain the same for both sets. This results in two $2^3 3^6$ factorials of possible product descriptions, i.e. a total number of 5,832 possibilities per attribute set. Obviously, market researchers may only have a fraction of this number of possible stimuli be evaluated by respondents.

A1 denotes the set with solely technical product attributes to describe the dimensions on which the product is evaluated by the customer. It contains attributes A to I of Table X (page 99) which are solely product-referent or technical product descriptions.

A2 is the type of attribute set which has one technical attribute replaced by a user-referent attribute, i.e. the attribute 'pointing device' is replaced with the attribute 'firm reputation'. Attribute levels of both sets contain metric, ordinal, and nominal types of scale values, i.e. monotonic and nonmonotonic attribute levels. Exchanging only two attributes with the same number of levels (three) and the same type of scale (nominal), and holding everything else constant, ensures that no other influences emanating from the attribute set on the evaluation of the product is confounded with a manipulation of the type of attributes used (i.e. except for influences from outside the attribute set, for example differences from random grouping of respondents).

A pretest of the importance of two user-referent attributes, 'firm reputation' expressed in a brand name, and the 'importance of what others think of a laptop' (concerning the stimulus description) for the respondent´s own decision, revealed that possible buyers do not regard referent others´ opinions as important in making a purchasing decision for laptop computers. However, firm reputation, i.e. what the user thinks about the source of the product, was rated as an important decision attribute [6]. This justifies the inclusion of 'firm reputation' as the attribute manipulation for testing the influence of type of attribute on value structure (i.e. part-worths) and prediction. It is, at the same time, a test of the assumption of independence from irrelevant attributes (IIA).

Type of Factorial Design.

The variable type of factorial design has two dimensions: fractional factorial number one (FF1) and fractional factorial number two (FF2). They differ in the specific

[6] The overall importance of referent others was rated lowest in influence on the decision (2.33 on a category rating scale from 1 to 5), with the next lowest overall importance rating of 3.26 for weight of a notebook computer. The firm reputation, expressed in its brand name, had an overall importance rating of 3.63. Thus, permutation is not with the least important attribute(s), as in the study of Reibstein, Bateson, and Boulding (1988, p. 275), but with the one having exactly the medium importance rank (5 out of 9). For details of the pretest see Appendix I.

fractional factorial used, but they do not differ in their confounding structure. The specifics of these fractional factorials are provided together with a discussion of derivation of the stimulus sets for the conjoint experiment in later sections of this study. This manipulation allows for estimation of the magnitude of influence exerted by the specific fractional factorial design on the estimation of value structure and on predictive capability. Specifically, it allows to partition error in estimates in those resulting from sparseness in the design of the stimuli, and in error from judgments of the respondents.

Type of Model.

The variable type of model has two dimensions: the self-explicated model (SE), i.e. part-worths or component values of attribute levels are obtained through direct elicitation methods as for instance through ratings, and the traditional conjoint model (TC) which derives part-worths for each individual based on his stated overall value judgments for a set of stimuli. The self-explicated model is specified with equation (E3.1.3) on page 81, the traditional conjoint model in its additive form is specified in equation (E3.1.4) on page 81, and the latter´s extension to a multilinear form is given by equation (E3.1.9) on page 83. The decision which traditional conjoint model to apply in this study is (partly) determined by the fractional factorial design layout and its respective confounding structure (limiting the number and types of interactions possible in the model, i.e. the upper bound), and by the empirical data which are used to test for the presence of particular interactions. Only after these estimations and tests can the appropriate traditional conjoint model form be determined. An additive, main-effects model constitutes the "lower bound" of traditional conjoint model form. In accord with Bateson, Reibstein, and Boulding (1987) it is agreed in this study that only the self-explicated model form establishes (convergent) validity (a full discussion of this issue is provided in section 2.6 of this study).

Actual codes for model forms used in the tables and figures of this study are as follows (please, refer to the letters in brackets of Table X on page 99):

iAxD: Interaction of attribute A (Performance / Speed; ordinal scale) with attribute D (Weight; ratio scale);

iBxD: Interaction of attribute B (Keyboard Size; nominal scale) with attribute D (Weight; ratio scale);

iCxD: Interaction of attribute C (Display Type; nominal scale) with attribute D (Weight; ratio scale);

All of these attribute interactions are substantively plausible, as an attribute with a nominal or ordinal scale interacts with an attribute that has a metric scale. In the pretest, only interaction iAxD was revealed as possibly necessary due to negative attribute correlations.

Segmentation Method.

Three (benefit) segmentation methods are examined in this study: a hierarchical cluster segmentation method (HIC), non-hierarchical hard clustering methods (NHC), and fuzzy clustering methods (FUC). These are a posteriori approaches to clustering.

In the traditional a priori two-stage segmentation approach subjects are clustered into segments on the basis of characteristic variables of the respondents, for example demographics, psychographics, and other distinguishing characteristics (potential covariates). The choice of a priori segmentation bases is a question of managerial judgment based on prior experience, theory, or objectives, and not merely a question of method performance, as the meaningfulness of segments is dependent on criteria such as reachability, substantiality, and actionability of the segments chosen, to name a few. After clustering, the conjoint model is estimated at the segment level, resulting in segment-level part-worth estimates. The results of this method are dependent on the goodness of the managerial hunch, as well as on the appropriateness of the selected variable(s) to covary with value attributions to products. This approach is not pursued further in this study.

In a hierarchical cluster segmentation method (HIC), first a traditional individual-level conjoint model is estimated. Then, subjects are clustered hierarchically either on the basis of their stated preferences, i.e. their overall value judgments for a profile, $U(x_1, x_2, \ldots, x_j, \ldots, x_m)$, the (stated) criterion variable in the equations, or on the basis of respective part-worths, i.e. benefits attributed to a number of attribute levels.[7] At the second stage, part-worths are re-estimated across respondents within each of the resulting segments.

In non-hierarchical hard clustering segmentation methods (NHC) a traditional individual-level conjoint model is estimated as with HIC. Then, subjects are clustered on the basis of stated preferences (or other attitudinal measure towards the product profile), or on the basis of respective part-worths, and then part-worths are re-estimated on the segment level. However, Wedel and Kistemaker (1989) have proposed an approach that estimates segments and optimizes segment performance using an exchange algorithm. An alternative approach proposed by Helsen and Green (1993) is to re-cluster using different k cluster seeds and choose the number of clusters

[7] One may also think about clustering on the basis of importances, but importances are derived from benefit attributions and are therefore only indirect measures of attribute (level) utilities. Thus, this possibility is not explored in this study.

that gives the best estimate on the performance measure. This is the approach pursued in this study.

In fuzzy clusterwise segmentation methods (FUC) fuzzy segments, i.e. segments in which subjects may have only partial membership, are estimated using an iterative weighted least squares method. The partitioning of subjects into clusters with partial membership forces partial membership values of subjects in different clusters to sum to unit value, which is not the case in Hagerty´s factor solution. The fuzzy c-means clustering method is applied, here (Bezdek 1981).

Dependent Variables

The impact of variations in conjoint methodology (type of attribute set, fractional factorial design, conjoint model form) and segmentation method (HIC, NHC, FUC) with according segment-level benefit estimation on surrogates for prediction of market choice (i.e. purchase likelihood) is assessed. Ancillary measures were collected as potential covariates and potential a priori segmentation bases. These were explained subsequent to product attributes for the traditional (i.e. the base) conjoint experiment.

When evaluating performance of conjoint models to measure customer value (structure) we want to choose those methods or procedures, and those models that are most reliable and valid with respect to specific managerial objectives. Unfortunately, as has been demonstrated in section 2.6, there is no such universal measure of overall "goodness-of-fit", reliability, or validity. Rather, different measures allow evaluation of performance from different perspectives, or for different purposes. This is an issue of relevancy of methods which cannot be answered objectively but only subjectively within the triangle dependencies of the research objective(s), i.e. the problem and its representation, the researcher, and the problem context, i.e. the environment or environmental conditions. Therefore, this study employs several (surrogate) performance measures for prediction of market preference and choice behavior to answer the research questions. They may be classified into absolute, incremental, and parsimonious fit and performance measures based on the calibration and/or holdout samples. The following paragraphs present these measures and provide rationales for inclusion in this study.

Absolute fit and performance measures determine the degree to which the overall model predicts the observations. These measures are most meaningful in comparison with those obtained through alternative models, or with additional information about the observations (e.g. together with standard deviations) that puts the measures´ magnitudes into perspective.

Root Mean Squared Error (RMSE).

The root mean squared error (RMSE) between stated and predicted purchase likelihood is calculated for a holdout sample of product profiles. Additionally, RMSE between stated and predicted purchase likelihood is calculated for the calibration sample of profiles as an internal consistency check (i.e. remaining magnitude of error or lack of fit of the conjoint model):

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{K}(Y_k - \hat{Y}_k)^2}{N}}$$

where K denotes the number of observations/predictions, i.e. profiles,

Y_k denotes the actual response,

$\hat{Y}_k$ denotes the prediction of Y_k , and

N denotes the number of respondents.

Purchase likelihood is a rating scale ranging from 0 (definitely would not buy this notebook computer) to 100 (definitely would buy this notebook computer). The RMSE is useful as all responses of an individual are on the same scale, and exhibit the same response pattern (e.g. "averager," or "extremist"). Though no threshold level may be established for "good" or "poor" remaining error per se, one may assess the practical significance of the magnitude of the RMSE when comparing it to the calibration sample and the magnitude of the scale (0 to 100 in this study). Details of elicitation of judgments are provided in the section about data collection and experimental procedures.

Pearson´s Product Moment Correlation Coefficient (r_{xy}).

The Pearson product moment correlation coefficient (r_{xy}) measures the strength of *linear* association between variables. It is calculated as:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}_i)^2 \sum_i (y_i - \bar{y}_i)^2}}$$, for all observations i and two samples X and Y.

Its property of being dimensionless allows for easy comparison across subjects. However, curved relationships between variables, no matter how strong, need not be reflected in the correlation. The same is true if data is clustered, and though the clusters show strong correlation within each. Also, r_{xy} and OLS regression are not resistant, i.e. influential observations or incorrectly entered data points can greatly change the measure. Therefore, correlations should be evaluated together with scatterplots of the calibration sample, as has been done in this study. Just as calculation often adds to the information provided by a scatterplot, a plot is essential if calculation is not to be blind and misleading. Examples of scatterplots of model observations and predictions, and associated performance measures are provided in Appendix V. A major problem with r_{xy} in most conjoint analyses remains the fact when the number of parameters is close to the number of profiles rated, r_{xy} will artificially inflate the correlation between observed and estimated scores of the calibration set of profiles due to overfitting. As this is no problem for the holdout sample, only those correlations are compared. Additionally, error degrees of freedom in this study are eleven (11) and nine (9) for individual-level models which should be enough to exclude distorting influences on r_{xy} via too few degrees of freedom for error. Another caveat is appropriate when correlations based on averages are applied to individuals: usually, these (average) correlations are too high. Finally, in tests that use the correlation coefficient r_{xy}, and those coefficients show non-marginal differences (i.e. high variation among coefficients), it may be a problem that r_{xy} is not interval-scaled. Therefore, in these cases, Fisher's z-transformation of r_{xy} is applied in order to transfer the scale of r_{xy} into an interval scale, except for at the extreme ends. Fisher's z-transformation is calculated as:

$$z (r_{xy}) = \frac{1}{2} \ln \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right)$$

First Choice Hit (First-Hit).

The first choice hit rate (First-Hit) is calculated as a percentage of correctly predicted choices for a holdout sample of sixteen (16) profiles, arranged into four (4) sets of four (4) product profiles per set:

$$\text{First-Hit} = \frac{\text{Count of Correctly Predicted Choices}}{\text{Number of Possible Choices}}$$

First-Hit is calculated because r_{xy} takes all choices into consideration, and in a sense, dilutes the direct relevance to the marketer of a product whereas the first choice hit rate is a more direct measure of market choice. After all, the typical customer will ultimately pick only one brand among the many available in the marketplace. For the prediction of First-Hit, value maximization is assumed as the choice rule, as opposed to rules like BTL or logit transforms[8].

Incremental fit and performance measures compare the performance of the proposed model to some baseline model, most often referred to as the null model. The null model should be some realistic model that all other models should be expected to exceed. In most cases, the null model is a single-construct model. In our case of prediction this is simply the overall mean without regard to any effects.

Coefficient of Determination (R^2).

This coefficient is calculated as follows:

$$R^2 = \frac{\text{Sum of Squares for Model (with Effects)}}{\text{Sum of Squares for Mean Model (w/o Effects)}}$$

R^2 between stated and predicted purchase likelihood is calculated for the calibration sample of profiles as an internal consistency check (i.e. goodness-of-fit of the conjoint model). R^2 estimates the proportion of variation in the response around the mean that can be attributed to terms in the model rather than to random error. It is also the square of the (Pearson product moment) correlation r_{xy} between actual and predicted response. For a derivation of this equivalence cp. Pedhazur 1982, esp. p. 21 and equation (2.18), or Moore and McCabe 1989, pp. 203.

[8] The BTL (Bradley-Terry-Luce) model computes the probability of choosing a profile as most preferred by dividing the profile's utility by the sum of all sample profile utilities. The logit model is similar to BTL but uses the natural logarithm of the utilities (SPSS, Inc. 1994, p. 32). Most applications of First-Hit, though, use value maximization (Wittink, Vriens, and Burhenne 1994, p. 47).

Parsimonious fit and performance measures relate some goodness-of-fit index of the model to the number of estimated coefficients required to achieve this level of fit. The basic objective is to diagnose whether model fit has been achieved by "overfitting" the data with too many coefficients. Their use, in most instances, is limited to comparisons among models, rather than to statements about substantive findings.

Adjusted Coefficient of Determination (Adj R^2).

This measure is calculated as:

$$\text{Adj } R^2 = 1 - \frac{\text{Mean Square for Model Error}}{\text{Mean Square for Mean Model}}$$

Adj R^2 adjusts R^2 to make it more comparable over models with different numbers of parameters by using the degrees of freedom in its computation of the mean squares. It is calculated between stated and predicted purchase likelihood for the calibration sample of profiles. With correlations for the holdout samples and appropriate transformations, R^2 and Adj R^2 measures can be compared for external validity. Despite the other performance measures available, R^2 is calculated to reveal the magnitude of an effect, here, by comparing differences in variance accounted for.

3.3.4 Construction of Stimuli

The construction of the stimulus profiles involves a number of preliminary considerations. First, the number of attributes and levels in Table X on page 99 allows for $2^3 3^6$ factorials of possible product descriptions, i.e. a total number of 5,832 possibilities per stimulus set. This number must be reduced to a set of profiles manageable for respondents. Green and Srinivasan (1978, p. 109) suggest an upper bound of about 30 profiles in commercial studies, and some more for student respondents (see also Louviere 1988, p. 58). Second, with nominal scales for attribute levels, it is desirable to estimate not only the main effects, but also at least selected two-way interactions (Louviere 1988, p. 58). The issue is whether predictive accuracy would be better with interactions because of increased realism or worse because of decreased degrees of freedom (DFs), increasing bias in estimation. Third, in order to obtain useful results and be able to perform desired tests, one replication is necessary in this study. Fourth, how can the evaluation task of the holdout sample of profiles be made easier after respondents performed the calibration task (to alleviate possible fatigue) ? Fifth, only an orthogonal design gives unbiased estimations (Johnston 1984, p. 172; Louviere 1988, p. 61), though corrections are possible (cp. Addelman 1962, Appendix B, Pedhazur 1982, pp. 371).

Given the objectives of the study to test for effects of two different fractional factorial designs on estimation of conjoint model on the individual level, and given above considerations, three different fractional factorials are necessary: FF1 and FF2 to test for the effect of the fractional factorial chosen, and one fractional factorial for the holdout profiles. FF1 and FF2 are two different Resolution IV fractional factorial designs with selected interactions obtained from Addelman (1962 ; also in Connor and Young 1961, p. 40 and Green, Carroll, and Carmone 1978).[9] Respondents evaluated 27 profiles of the $2^3 3^6$ factorial. The estimation of main effects (not confounded with two-way interactions, but with higher ones assumed to be zero) uses up sixteen (16) degrees of freedom (DFs; incl. the intercept), leaving eleven (11) for selected two-way interactions. Given Johnson, Meyer, and Ghose´s (1989) finding that negative correlations between attributes (in contrast to positive correlations) might pose a problem for conjoint experiments, and given the finding of the pretest that only one two-way attribute correlation (between weight and performance) was negative, the

[9] A more detailed discussion of obtaining the two fractional factorials is provided in Appendix IV/4.

assignment of attributes to design columns was chosen so that this possible interaction between weight and performance could be tested, as well as selected positive ones. For the holdout sample of 16 profiles, an easier to evaluate Resolution III (extreme) design (in the three-level case) was chosen including only the two (2) extreme levels per attribute. The respective coding structure for FF1, FF2, and the holdout (HF) is provided in Appendix IV.

3.3.5 Data Collection

Sample

Subjects for this study were sampled from the business school of a medium-size Northwestern university. Participants were undergraduate and graduate students out of seven (7) different classes; six (6) in marketing and one (1) in organizational behavior with marketing topics. This should have yielded somewhat homogeneous respondents with respect to the measurement environment, at least in their pursuit of educational achievement and possibly in their attitude towards surveys, and their state-of-mind towards the measurement object (i.e. laptop computers; for an examination of state-of-mind effects on the accuracy of value measurement cp. Wright and Kriewall 1980). Final sample size reached 117 useful responses on a voluntary basis. Some respondents had to be deleted because of missing cells or inability or unwillingness to (completely) perform the task (DeSarbo, Wedel, Vriens, and Ramaswamy 1992, p. 284 found only 2 respondents out of 48 to be unable or unwilling to perform the task). This yielded in between 432 (16 profiles x 27 respondents for group G3) to 480 (16 profiles x 30 respondents for the other groups) observations per group as the basis for calculating group (average) performance measures R^2, Adj R^2 , RMSE, and r_{xy} , and it resulted in 108 (4 x 27) to

120 (4 x 30) observations for First-Hit (cp. timely design of administration of data collection in Figure 8 on page 92).

Students are considered to be an appropriate target population for this study. They are usually at least somewhat familiar with computers, and may be considered among the target population of buyers of notebook computers. Virtually every notebook computer manufacturer for the consumer and small business market provides educational discounts or other financial incentives for student buyers. Additionally, students may be considered reasonably interested in the product class to carefully conduct demanding data collection procedures without adequate financial compensation. In particular, a convenience sample of students is justifiable, given that the purpose of this study is to investigate effects of simultaneous methodological variations and subject grouping without necessarily generalizations to a larger population. Finally, in a number of conjoint studies with student samples and samples taken from a different target population at a later time, no serious unexpected negative or contradictory effects are reported.[10]

Administration

An overview of the design layout concerning administration of measurement is provided in Figure 8 on page 92. Respective groups and observations as they are identified in the results section are provided in brackets in the subsequent presentation. Two sessions were conducted with each individual which necessitated recording of an identification variable, the student´s name and class number. For the first session (1st replication), subjects were randomly assigned to two groups, R1 (G1) and R2 (G2),

[10] While comparisons between these two populations never changed substantive findings, on average, i.e. over all the treatments, students showed generally higher reliability than representative samples of the population (Reibstein, Bateson, and Boulding 1988, p. 284).

which differ in the type of attribute set they evaluated (A1 or A2). Both groups, however, evaluated the attribute sets on the same fractional factorial design (FF1). For each group four types of observations were recorded:

(1) Self-explicated desirability ratings of attribute levels, and ratings of attribute importances anchored with respective best attribute levels (OS; for a rationale for anchoring of importances cp. Srinivasan 1988, p. 296).

(2) Conjoint ratings of product stimuli (OC).

(3) Recording of demographic variables (OD).

(4) Holdout sample ratings of product profiles and first choice out of four (4) sets of four (4) stimuli per set (OH).

In the second replication, subjects of the former groups R1 and R2 (G1 and G2) were again randomly assigned to two further groups within the first group assignment, resulting in a total of four groups (R11, R12, R21, and R22; identified as groups G3, G4, G5, and G6 in the results section of this study). The treatments now varied in the type of fractional factorial used (FF1 or FF2) for the first two groups (R11, R12; respectively G3, G4), and in the type of attribute set (A1, A2) for the second two groups (R21, R22; respectively G5, G6). This arrangement is necessary in order to isolate effects of reliability over time from effects of reliability over attribute set, and reliability over stimulus set without resorting to solely between-subjects comparisons. Details of the analyses are provided in the subsequent section. For each group, all four (4) types of observations of the first replication were also recorded in the 2nd replication. Though only three types of observations in the 2nd replication are needed for the analyses (OS, OC, OH), the additional recording of demographics allows for reliability checks of responses which should not differ from the first responses, and which are usually assumed to be very reliable over a variety of measurement conditions. Data, again, was collected with a self-administered questionnaire. The experiment was confidential though not anonymous, i.e. student´s name and class number served as the matching code for repeated measurements, which was later recoded into a unique respondent number (SID).

3.3.6 Analysis

The main objective of this study is to test relative influences of

- selected methodological variations of conjoint analysis and
- segmentation methods (i.e. grouping of subjects)

on customer value structure, and in particular concerning changes in predictive accuracy. This is accomplished by testing hypotheses suggested by the literature review and accompanying research questions (section 3.2 on page 86). The hypotheses pertaining to the first four research questions test influences of type of attribute and factorial design on predictive accuracy for individual models, as well as their relative performances. Hypotheses pertaining to research questions number five (5) and six (6) test relative influences of different conjoint models on predictive performance for segment-level models, and research questions number seven (7) and eight (8) do not lend themselves to hypothesis testing but are subject to interpretation of test results in prior stages of this study.

Phases I and II

For the following discussion, please refer to the overview of study objects in Figure 9 on page 118. In Phase I of the analysis, individual-level multiattribute preference models were estimated for self-explicated level desirabilities and importance ratings, and the conjoint task, based on OS and OC, respectively. This yielded six (6) different groups of SE and TC models with exposure to different methodological factors (attribute set, fractional factorial, and time), and with respective part-worth utilities for each individual (i.e. individual value structure), subsequently denoted as PW1 to PW6. In addition, within each group, four (4) different TC models were estimated, one (1) main effects model, and three (3) different models with one two-way interaction, resulting in a total of five (5) different preference models per individual in a group. Therefore, a total of 1170 individual preference models were estimated in this phase (i.e. 117 respondents x 2 replications x 5 models). Accordingly, overall utilities were predicted and performance measures were calculated, as for instance first choice hit rates, with these models using the stimulus profiles of the holdout samples (OH). For the TC models, R^2 and adjusted R^2 were calculated for the calibration profiles, yielding R^2 and adjusted R^2 for all six calibration groups (i.e. TC-$R^2$1 to TC-$R^2$6, averaged over the individuals in the group). Accordingly, R^2 , Pearson´s product moment correlation coefficient r_{xy} , Fisher´s z-transformation of r_{xy} , first-choice hit rates (First-Hit), and root mean squared error (RMSE) for the holdout profiles were calculated for all individual-level models. RMSEs were also calculated for TC calibration profiles.

For Phase II, the segment-level analyses, part-worth utilities derived in Phase I for each individual by the overall best predictive TC model served as the inputs for the benefit segmentation methods (HIC, NHC, FUC), yielding segments based on benefit attributions to attribute levels. Then, part-worths were re-estimated for all segments of all three types of segmentation methods with the conjoint model form of the input TC model, predictions of the holdouts were performed with these segment-level conjoint models, and performance measures were calculated for all three (segmentation) types of models and all (117) observations in their respective three (3) or four (4) segments, denoted for instance as TAT-$R^2$1 to FUC-$R^2$4, or TAT-First-Hit1 to FUC-First-Hit4. This yielded six values per performance measure in Phase I (after averaging and weighting them over individuals for a particular model) in respective cells in Table VIII on page 90, and up to fourteen (14) values per performance measure (7 cluster models x 2 different numbers of clusters) in Table IX on page 91. An overview over study objects is provided in Figure 9 on page 118.

General Testing Procedures

Differences in R^2 and Adj R^2 for different types of conjoint models cannot be tested (though R^2 index can be tested for significance on its own, cp. Pedhazur 1982, pp. 57), but their magnitudes, their goodness-of-fits are evaluated according to guidelines provided in the Monté Carlo study of Umesh and Mishra (1990; goodness-of-fit, significance, and power are design-dependent). Differences in the other three (3) performance measures illustrated in section 3.3.3 (pp. 106; RMSE, Fisher´s z-transformation of r_{xy}, and First-Hit) can be tested for significance with different testing procedures.

With a little modification, these three measures can be tested for with one-way ANOVA, testing the hypothesis that two sample means μ_1 and μ_2 are indifferent:

H_0: $\mu_1 = \mu_2$; H_a: $\mu_1 \neq \mu_2$.

Groups | **Model Types** | **Measures**

Phase I

Label	Size
R1 (G1)	57
R2 (G2)	60
R11 (G3)	27
R12 (G4)	30
R21 (G5)	30
R22 (G6)	30

	Benefits	Performance	
	(Value Structure)	(calib.)	(hold.)
SE	• unscaled part-worths, • scaled part-worths (positive interval; PW1, PW2, …, PW6) [a], • normed importances (Imp1, Imp2, …, Imp6)	n/a	• RMSE, • r_{xy} , • z (r_{xy}), • R^2, • First-Hit, • First-Hit (mean counts)
TC main effects (1x)	• unscaled part-worths, • scaled part-worths (positive and negative intervals; PW1, PW2, …, PW6) [a], • normed importances (Imp1, Imp2, …, Imp6)	• R^2 , • RMSE, • Adj R^2	• RMSE, • r_{xy} , • z (r_{xy}), • R^2, • First-Hit, • First-Hit (mean counts)
TC w/interaction (3x)	…	…	…

Phase II

Choose best TC model; segment; compile groups

Phase II

Choose best TC model; segment; compile groups

3-Cluster Segments (c1 to c3)	0 to 117
4-Cluster Segments (c1 to c4)	0 to 117

HIC	• unscaled part-worths, • scaled part-worths (positive and negative intervals; PW1, PW2, ..., PW6), • normed importances (Imp1, Imp2, ..., Imp6)	• R^2, • RMSE, • Adj R^2	• RMSE, • r_{xy}, • z (r_{xy}), • R^2, • First-Hit, • First-Hit (mean counts)
NHC	...	...	...
FUC	...	...	...

(calib.) = calibration set

(hold.) = holdout set

a = These two scaled (as well as the unscaled) part-worths are not comparable across types of models, i.e. SE and TC.

Figure 9. Study Objects.

In order to test differences between two independent groups (two-group univariate analysis), the better test is a t-statistic (a special case of ANOVA), though an F-statistic is more common[11].

In order to test for differences among k independent groups, the appropriate test statistic is the F-statistic resulting from ANOVA. These test statistics are only (at least formally) valid if their assumptions are met, i.e. if the dependent variable is normally distributed, and if variances are equal across groups. However, there is evidence that F-tests in ANOVA are quite robust with regard to violations of these assumptions. But these F-tests are sensitive to outliers and their impact on Type I error (Hair, Anderson, Tatham, and Black 1992, p. 159 with additional references; p. 160)[12]. Due to the considerable time lag between the two measurements of each individual (between two and four weeks), memory effects may reasonably be assumed to be negligible, given the variations and difficulty of the evaluation task (Reibstein, Bateson, and Boulding 1988; McCullough and Best 1979, who measured reliability over time after two days, and with only three attributes in the profiles). Therefore, tests do not include repeated measures ANOVAs. However, paired t-tests are performed in Phase I and their results are compared to the F-tests in order to separate

[11] The t-statistic in this special case is preferred to the F-statistic because of its greater robustness concerning violations of assumptions, i.e. deviations from normality and skewedness of the distributions, as well as unequal group variances, though the ANOVA F-statistic is also quite robust (Moore and McCabe 1989, p. 520, pp. 546, p. 565, pp. 568, and pp. 721; extensive simulations may be found in Posten 1978). Other arguments for prefering the t-ratio over F are provided in Pedhazur (1982, pp. 28), the most important of which is ability to calculate confidence intervals. This ANOVA t-statistic is equivalent to an F-statistic with 1 DF in the numerator, and a pooled error variance ($t^2 = F$; cp. Pedhazur 1982, p. 28). A paired t-test, however, uses the separate standard errors of the two groups and is therefore able to provide a more accurate test.

[12] Moore and McCabe (1989, p. 721) argue against formal tests for equality of variances, as these tests suffer from similar deficiencies as those deficiencies they are testing for. Instead, they suggest a general rule of thumb to compare the ratio of the largest group (sample) standard deviation to the smallest group (sample) standard deviation. If this ratio is less than two (2) "the results will still be approximately correct" (p. 722).

within-subject effects from between-subjects effects. Post hoc tests for multigroup ANOVAs to pinpoint exactly where significant differences lie were not of interest in this study. However, multigroup ANOVA tests were conducted in order to determine significance of differences in performance among segmentation methods.

RMSEs can be used directly for significance tests concerning group differences on methodological variations. r_{xy} can also be used directly if individual-level correlation coefficients are used, as is the case in this study, and these are tested on their means. But, as paired t-tests are also used to separate individual from group differences which necessitates an interval scale (i.e. with calculation of differences between scale values), and in order to compare both types of tests, Fisher's z-transformation of the correlations of r_{xy} are used for both types of tests. For First-Hit to be tested with ANOVA, the test has to be performed on the mean counts of first choices of the groups (cp. Green, Helsen, and Shandler 1988). For r_{xy} and First-Hit there are also more direct tests available.

For r_{xy} a two-tailed or one-tailed (if the direction of the difference can be hypothesized in advance) two-sample z-test of significance may be conducted with $z(r_{xy})$ values as data (cp. Boecker and Schweikl 1986, pp. 22 or Yamane 1973)[13]. The test statistic for two samples 1 and 2 is given as (Schaich 1977, pp. 209):

[13] Fisher's z-transformation on r_{xy} is theoretically necessary because the raw r_{xy}s are not interval-scaled (Bortz 1979, pp. 260; Hartung, Elpelt, and Klösener 1984, p. 549). However, Green, Helsen, and Shandler (1988) report replications of their ANOVAs with Fisher's z-transformation without changes in the substantive findings, adding to the notion that ANOVAs are quite robust with respect to violations of assumptions (footnote 4 on p. 395).

$$z = \frac{\bar{z}(r_{xy_1}) - \bar{z}(r_{xy_2})}{\sqrt{\frac{1}{n_1(K_1-3)} + \frac{1}{n_2(K_2-3)}}}$$

where K is the number of objects′ pairs the correlation index is based upon, and

n denotes the number of correlation coefficients.

For First-Hit, a z-test on proportions may be used with the following statistic (Schaich 1977, p. 213):

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

where p denotes the respective proportion, and

n is the number of responses on which the proportion is based.

A more common practice for a z-test on proportions under H_0: $p_1 = p_2$, but with one more calculation (i.e. pooling of proportions), is provided with (Moore and McCabe 1989, pp. 597):

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}_1, \hat{p}_2$ denote respective sample proportions, and

$\hat{p}$ denotes the pooled estimate of p, i.e. the overall proportion of successes in both samples.

Both z-tests for First-Hit are based on the normal approximation to the binomial distribution. Thus, as a general rule, this is valid when n_1p_1, $n_1(1-p_1)$, n_2p_2, $n_2(1-$

p_2), $n_1\hat{p}$, $n_1(1-\hat{p}$), $n_2\hat{p}$, and $n_2(1-\hat{p}$) are all greater than 5 (Moore and McCabe 1989, p. 596 and p. 598). Though these z-tests can be more powerful, they are also more sensitive to deviations from the underlying assumption of normal distributions. For a two-group univariate ANOVA with unequal cell sizes *and* small cells (less than thirty observations per group), which may occur only for group R11 (G3), the t-statistic is adjusted to a t-distribution as proposed by Moore and McCabe (1989, p. 541 and pp. 546; also Hines and Montgomery 1990, p. 304 and pp. 310; known as Behrens-Fisher problem)[14]. The last three more direct tests for r_{xy} and First-Hit did not yield substantial differences to the F-tests and paired t-tests performed, though in general they indicated slightly higher significance, i.e. lower p-values for treatment effects. Therefore, they are not reported, here. The specific tests performed to answer the research questions are detailed in the last paragraphs of this section.

As there is considerable confusion as to the proper application and meaning of tests in conjoint analysis, some remarks about their use in this study seem appropriate. It is well known that each test of significance is valid only in certain circumstances and for specific assumptions, with properly produced data being particularly important. Concerning the problem of choosing a level of significance, there is no sharp border between "significant" and "insignificant," only increasingly strong evidence as the P-value decreases (cp. Moore and McCabe 1989, pp. 485; Pedhazur 1982, p. 24). There is no practical distinction between the P-values 0.047 and 0.051, and it makes

[14] The two-sample t-statistic so common in ANOVA does not have a t-distribution because a t-distribution replaces a N(0,1) distribution only when a single standard deviation in a z-statistic is replaced by a standard error. Here, two standard deviations are replaced by the corresponding standard errors, which does not produce a statistic having a t-distribution. This, however, can be remedied with appropriate adjustments, namely (1) with an approximation to a t-distribution by adjusting the DFs from sample data, or (2) by taking a t^* of the smaller group size (n-1) which leads to a more conservative test (cp. Moore and McCabe 1989, p. 541).

no sense to treat $\alpha = 0.05$ as a universal rule for what is significant[15], as it is also well known that given a sufficiently large sample, the probability of rejecting the null hypothesis is high. Rather, the researcher may first decide upon the magnitude of the effect, or relation, magnitudes of differences between means, magnitudes of treatment effects, and the like that is to be considered substantively meaningful in a specific area of research. Then, the level of significance (Type I error) and the desired power of the statistical test (1 – Type II error) are selected. Often, however, even those "meaningful" magnitudes cannot be determined adequately in advance (i.e. though studies may have found significant differences, the spread around the parameter, the confidence interval, may be too large to allow for precise figures, or the figure may be so small relative to other influences, that it may not be of practical relevance, cp. Moore and McCabe 1989, p. 544). Moreover, meaningfulness is specific to a given research area, and there are no generally applicable and no objective criteria for meaningfulness of findings (for an extensive discussion of this topic, see Cohen 1977). Finally, researchers in the same research area may disagree about the meaningfulness of a finding when they consider, for instance, the costs involved in obtaining it, or when they consider mean differences of groups in light of individual variances.

In conjoint analysis, and one may conjecture in wholistic judgments of product preference in general, the emphasis of analyses is not on tests of significance, but on estimation of parameters which, admittedly, is not independent from each other. Nevertheless, the analogy of judgments about the significance of effects in conjoint analysis to judgments about the inclusion of additional parameters in stepwise regression is also not quite appropriate. In stepwise regression, parameters are added or removed depending on significance tests of adding or removing variables to the

[15] For a conjecture why a significance level of $\alpha = 0.05$ is so universally accepted in science cp. Moore and McCabe 1989, pp. 486.

regression equation. However, in studies involving product preference and choice behavior, removing or only altering unimportant attributes or insignificant parameters changes the decision context, and was found to lower predictive accuracy (Green and Schaffer 1991). The counter-intuitive finding that non-significant parameters (associated with unimportant attributes) contribute to a significant improvement in prediction[16] suggests "unimportant" does not mean you are allowed to neglect or ignore the parameter because of lack of significance or, conversely, because of lack of power. Rather, there are two options as remedies for attaching too much importance to statistical significance in this situation:

(1) Increase the significance level α to a level higher than 0.05; an increasing number of studies use $\alpha = 0.1$, and/or report actual P-values (e.g. Reibstein, Bateson, and Boulding 1988, Table 2 on p. 281; Ostrom and Iacobucci 1995, Table 2 and subsequent discussion) rather than reporting just significance with special characters;

(2) Plot the data, examine them carefully, and report results with confidence intervals, as a confidence interval actually estimates the size of an effect, rather than simply asking if it is too large to reasonably occur by chance alone.

The first option can be done without methodological problems, while the second option, estimation of confidence intervals, assumes equal variance among observations, which may not always be a valid assumption in this study. Nevertheless, both options are applied here.

[16] Note that significance is here at two different levels, at the parameter level, and at the criterion level, i.e. the whole model.

Hypotheses and Associated Tests

Research Question # 1.

What is the influence of the type of attribute chosen for the evaluative task (i.e. technical or product-referent attributes and non-technical or user-referent attributes) on customer value structure and predictive validity ?

From the literature review, hypotheses for the attribute set may be stated as follows:

H_0: The inclusion of user-referent attributes in the attribute set does not increase predictive performance.

H_A: The inclusion of user-referent attributes in the attribute set does increase predictive performance.

The hypothesis is tested with one-way ANOVA and paired t-tests in the form of between-subjects and within-subjects group comparisons. This can be accomplished in two ways (cp. Figure 8 on page 92). First, differences in prediction for groups R1 (G1) and R2 (G2) can be tested. Second, differences in prediction due to the attribute set can be calculated by comparing group observations five and six (e.g. SE-RMSE5 and SE-RMSE6) for all five types of models (SE, TC main effects, 3 models TC with interactions). A repeated measures design allows for the isolation of error due solely to the individual (i.e. measurements at different times), and error due to the treatment effect (i.e. different attribute sets) leading to increased precision in the analysis (Pedhazur 1982, p. 559). Therefore, the first comparison between groups R1 (G1) and R2 (G2) is not performed. However, the comparison between R21 (G5) and R22 (G6) is confounded with variations due to the time of administration. Therefore, the variation due to time of administration is computed for the same group of individuals between observations two and five in order to gauge the variability due solely to time of administration. Additionally, the variation due to the confounded effects of time and variation in the attribute set is calculated for the same group of individuals between observations two and six. While the former test constitutes a test of reliability over time, the latter constitutes a test of reliability over attribute set. For increased clarity, Table XI on page 131 at the end of this section provides an overview over group comparisons for performing tests of hypotheses for Research Question # 1 to Research Question # 3 with respective rationales.

Research Question # 2.

> What is the influence of specific factorial designs, i.e. of specific combinations of product attribute values, on estimation of customer value structure and predictive accuracy ?

From the literature review, hypotheses for the stimulus set may be stated as follows:

H_0: The utilization of a specific fractional factorial design does not influence predictive performance.

H_A: The utilization of a specific fractional factorial design does influence predictive performance.

This hypothesis is also tested with one-way ANOVA and paired t-tests in the form of between-subjects and within-subjects group comparisons. However, this can only be accomplished in one way (cp. Figure 8 on page 92 and Table XI on page 131). Differences in prediction due to the fractional factorial design are calculated for the same group of individuals by comparing group observations three and four (e.g. SE-

RMSE3 and SE-RMSE4) for all five types of models. This comparison, again, is confounded with variations due to the time of administration. Therefore, variation due to time of administration is computed between observations one (Part R1 or G1) and three (R11 or G3) in order to gauge the variability due solely to variation in the different sets of fractional factorials used for construction of the stimulus profiles. This constitutes a test of reliability over time, while the former test constitutes a test of reliability over stimulus set.

Research Question # 3 .

How do type of attribute in the product profile and factorial design interact in their influence on customer value structure for different models ?

From the literature review, no indication about the direction of this interaction for predictive accuracy is obtained. One general suggestion is that differences due to several methodological variations should cancel out. This leads to the following hypothesis:

H_0: The interaction of differences in attribute set and specific fractional factorial design does not influence predictive performance.

H_A: The interaction of differences in attribute set and specific fractional factorial design does influence predictive performance.

This hypothesis is tested using one-way ANOVA for performance measures and the five types of models. However, in this case paired t-tests cannot be employed to isolate effects due to time from effects of treatment interactions. Observations for group four (R12 or G4) and group five (R21 or G5) are utilized to test this hypothesis (e.g. SE-RMSE4 and SE-RMSE5). A test with observations for group four (R12 or G4) and a subsample of observations for group two (R2 or G2), however, will not allow for isolation of effects due to time and group assignment. If the former (between-subjects) test can be interpreted as a test for the interaction effect of attribute set and factorial design variations, or if time and random group assignment may be causes of possible deviations depends on the outcomes of tests pertaining to research questions number 1 and number 2.

Research Question # 4.

Which individual-level model for customer value structure performs best with respect to prediction ?

From the literature review, the only indication about the direction of relative performance of individual-level models is suggested superiority of (traditional; TC) conjoint models over self-explicated (SE) models. However, for methodological variations and a variety of situations no general statements about predictive accuracy of models with interactions and without them was obtained. This leads to the following hypothesis:

H_0: Individual-level models for customer value structure do not distinguish themselves in terms of predictive performance.

H_A: Individual-level models for customer value structure do distinguish themselves in terms of predictive performance.

This hypothesis is tested using multi-way ANOVAs for performance measures and the five types of models. The tests are performed with all 2nd group estimates and selected performance measures (Fisher´s z-transformed correlation coefficients, RMSE, and First-Hit).

Research Question # 5.

Can cluster-based segmentation approaches improve accuracy in prediction of value attributions to product profiles over individual-level conjoint models ?

Research Question # 6.

Which aggregate model for customer value structure performs best with respect to prediction ?

From the literature review about the nature of value, specifically its conceptualization/ representation as a ratio between (perceived) benefits and sacrifices, in section 2.1.1 (pp. 29) of this study it was concluded one may reasonably well assume highly idiosyncratic sets of relevant attributes and model forms. This also suggests that individual-level conjoint models should outperform segment-based conjoint models in terms of predictive accuracy. However, more recent literature and pilot studies about aggregate conjoint models suggests that segment-level based methods should outperform individual-level part-worth utility models because of more stable parameter estimates, though there may be increased individual variance. This claim has not been confirmed in one replication of one particular model. Therefore, the hypothesis for this research question may be stated as follows:

H_0: Segment-level part-worth utility models do not influence predictive performance.

H_A: Segment-level part-worth utility models do influence predictive performance.

This hypothesis is tested by performing one-way ANOVA on selected pairs of segment-level models and over selected performance measures. In order to compare segment-level and individual-level models, and to address violations of test assumptions, paired t-tests and Chi-Square tests are conducted for the segment-level comparisons.

Research Question # 7.

> Are the purposes of prediction and segmentation, as well as potential other purposes, better served with the suggested methods, and what practical limitations are there for the different methods to support specific purposes ?

Research Question # 8.

> Are benefit segments obtained with different clustering procedures meaningful for target marketing, or may they only increase predictive accuracy ?

These two questions do not lend themselves to hypothesis testing. They concern the benefit cluster solutions obtained, and possible conflicts from high predictive accuracy but poor ways to meaningfully address segments with various business policies.

TABLE XI

OVERVIEW OVER COMPARISONS OF GROUPS FOR TESTS OF HYPOTHESES OF RESEARCH QUESTIONS NUMBER ONE (# 1) TO THREE (# 3)

	Group Comparisons		Comments
	• Between-subjects • Same time; 2nd administration	• Within-subjects • Different times; 1st and 2nd administration	
	• One-way ANOVA	• One-way ANOVA • Paired t-tests	Tests applied in this study; more specialized tests yielded only minor deviations to the paired t-tests
# 1:	Attribute Set (A1, A2)		
	G5 <—> G6		Not G1 <—> G2 because comparison does not allow for the isolation of error due solely to the individual
			Not G3 <—> G5 because effects of time, i.e. reliability, cannot be isolated and compared with the confounded effect
		Part G2 <—> G5 (one-tailed, $\alpha = .1$)	Difference solely due to time of administration
		Part G2 <—> G6 (two-tailed, $\alpha = .1$)	Difference due to confounded effects of time of administration and attribute sets
# 2:	Factorial Set (FF1, FF2)		
	G3 <—> G4		Not G4 <—> G6 because effects of time, i.e. reliability, cannot be isolated and compared with the confounded effect
		Part G1 <—> G3 (one-tailed, $\alpha = .1$)	Difference solely due to time of administration
		Part G1 <—> G4 (one-tailed, $\alpha = .1$)	Difference due to confounded effects of time of administration and factorial sets
# 3:	Interaction Attribute / Factorial Sets		
	G4 <—> G5		Sole interaction, but different first administration
			Not part G2 <—> G4, as difference due to confounded effects of time of administration and attribute / factorial sets interaction

CHAPTER IV

RESULTS

This chapter presents results for Phases I and II of this study, i.e. answers as they pertain to research questions. First, some preliminary remarks about how to achieve comparability for different types of models in terms of value structure and performance measures are provided. Second, results of Phase I, i.e. the individual-level analyses, are provided. Third, segment-level analyses are presented as obtained in Phase II. Finally, results are summarized and interpreted in the following chapter.

4.1 Comparability, Research Strategy, and Individual Reliability

In order to ensure "fair" comparison among study objects (cp. Figure 9 on page 118), inputs, method, and output had to be adjusted for most of the analyses. Choices were guided by two objectives: Let the best of a method come to bear, and stay closest to the original data. This involves especially the transformation of derived and original part-worths into scaled part-worth utilities. Scaling of part-worths across subjects in individual-level conjoint analysis, as e.g. with normalization, is not appropriate because of changes in the relative contribution of attribute levels to overall preference, and because of response pattern influences which should somehow be preserved as information about respondents. Scaling for comparability is necessary, however, because the sum of the part-worth ranges (for computing importances, for instance) is

- a function of the number of parameters estimated with the model; the more means fitted, the higher the sum of the part-worth ranges, i.e. the lower the importance of a specific attribute. This is especially important in models with interaction terms.
- a function of the response pattern of an individual, i.e. "extremists" have large ranges, "equalizers" show narrow ranges among attribute levels.
- a function of possibly other systematic and random influences.

Therefore, tables of raw regression coefficients are not replicated, here, as their meaning is hard to interpret. Instead, value structure is presented with scaled part-worths and associated attribute importances for treatment groups. Signed utility levels are preferred to utilities scaled with offsets as the former provide information about positive or negative overall contribution, i.e. about magnitude and direction of change[17]. For self-explicated models (SE), there are only positive part-worths. Thus, scaled part-worths of SE-models are not directly comparable to those obtained with conjoint models. For this reason, value structure of SE models is presented separately as an overview over attribute importances for respective treatment groups in Table XIII on page 137. Also, prediction with SE-models were performed using the original responses, as studies found them to work better than scaled ones (Green and Schaffer 1991, p. 479).

[17] The scaling formula is provided as equation (E5.1) on page 199. Scaled part-worth utilities are to be interpreted as follows: From a general level of utility (least squares mean; intercept), given the product/attribute description, how much utility/disutility does a specific level have ? Large ranges of attribute levels may also be interpreted as exhibiting distinct preference structure.

Testing and research strategy used the following guidelines. Isolation of differences solely due to individuals (true error; over time) from differences due to treatments are only possible with repeated measures designs. Without them, i.e. leaving individual differences uncontrolled, they comprise part of the error term. In this study, due to the complexity of the task and the time between measurements (see section 3.3.6 for details), data may reasonably be assumed to be independent, allowing valid F-tests[18]. However, where possible, i.e. when study design and measure allowed it, paired t-tests were performed in addition to the F-tests, allowing for greater precision and confidence in the analysis on the level of the group. All data sets were checked for outliers with an outlier box plot, and a Shapiro-Wilk W test for the assumption of normal distribution of input data was performed. Where necessary, equality of variances was also checked for.

Considering the individual-level analysis which forms the basis for comparison of treatments, in many cases the effects of single predictors (i.e. attribute levels) in the individual model were not significant, however, the total model mostly was. In accordance with a majority of the conjoint and social science literature, individuals are considered reliable at a level of $\alpha \leq 0.1$. Table XII on page 136 provides an overview of individual reliability for different conjoint model forms, and over the measurement groups. Only few respondents showing non-significance of the model were identifiable as outliers when performing an outlier box plot on the group[19]. Outliers were not only observable on the low end, but on the high end as well, though even less so. For the first measurement (in time) and considering only main effects models, the

[18] When residuals are correlated due to repeated measures on the same subjects, which may usually be assumed, the F-ratio is only valid if stringent assumptions are met (details in Pedhazur 1982, p. 554).

[19] Outliers may be considered points outside the interval [lower quartile - 1.5*(interquartile range); upper quartile + 1.5*(interquartile range)] (cp. SAS Institute 1994a, pp. 34)

minimum (calibration set) R^2 of the individual model was 0.407, and maximum R^2 was 0.976. For the second administration, minimum R^2 was 0.481, and maximum R^2 was 0.990. These figures show poor fit for the minimum R^2 which explains only 40.7 or 48.1 percent of the variance. On the other hand, some respondents showed near perfect fit. Concerning the holdout set of profiles, respective figures are no fit (0.000) for the minimum and 0.983 for the maximum R^2 in the first administration, and 0.003 and 0.961 in the second administration which show extraordinarily high maximum values for some respondents, considering the complexity of the task. Average figures for all measures are provided in respective tables for tests of hypotheses.

TABLE XII

INDIVIDUAL RELIABILITY OVERVIEW (CONJOINT MODEL FIT AT P-VALUE < 0.1)

	Type of Model	Index	Measurement Group Reliability Indices							
			Part G1 (G3)	G3	Part G1 (G4)	G4	Part G2 (G5)	G5	Part G2 (G6)	G6
1.	TC main effects	Count	22	26	28	25	28	26	26	25
		Percent	81.48%	96.30%	93.33%	83.33%	93.33%	86.67%	86.67%	83.33%
2.	TC iAxD	Count	20	24	27	24	27	26	24	25
		Percent	74.07%	88.89%	90.00%	80.00%	90.00%	86.67%	80.00%	83.33%
3.	TC iBxD	Count	20	24	28	24	27	26	23	24
		Percent	74.07%	88.89%	93.33%	80.00%	90.00%	86.67%	76.67%	80.00%
4.	TC iCxD	Count	22	23	27	25	26	27	23	21
		Percent	81.48%	85.19%	90.00%	83.33%	86.67%	90.00%	76.67%	70.00%
5.	SE	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Partial groups represent the first measurement of the second column, or 2nd group of individuals (cp. Figure 8 on page 92).

TABLE XIII

ATTRIBUTE IMPORTANCES FOR SELF-EXPLICATED (SE) MODEL TYPE (ALL GROUPS)

Attributes:	Part G1 (G3)	G3	Part G1 (G4)	G4	Part G2 (G5)	G5	Part G2 (G6)	G6
Weight	11.41%	10.15%	9.81%	10.10%	8.59%	9.15%	10.26%	8.95%
ScrSiz	10.73%	10.20%	8.92%	9.92%	10.20%	9.17%	10.70%	11.05%
Display	11.83%	11.54%	9.93%	10.15%	11.55%	11.81%	12.08%	12.05%
B_Price	12.93%	13.51%	12.35%	13.19%	13.03%	13.95%	13.04%	13.02%
Keyb_Siz	9.43%	9.95%	9.26%	9.36%	9.53%	9.15%	9.21%	9.73%
BattLife	10.70%	10.87%	12.36%	11.71%	10.80%	11.08%	10.57%	10.86%
Speed	12.25%	12.38%	14.80%	12.15%	13.59%	13.20%	12.97%	12.87%
Features	12.17%	11.86%	12.85%	14.56%	12.41%	12.72%	11.53%	12.50%
PointDev[a)] / FirmRep[b)]	8.55% a)	9.53% a)	9.72% a)	8.88% a)	10.31% b)	9.77% b)	9.64% b)	8.97% a)

Partial groups represent the first measurement of the second column, or 2nd group of individuals (cp. Figure 8 on page 92).

4.2 Phase I

Presentation of results for tests of hypotheses are provided along the following lines. First, performance measures for methodological variations are presented. Next, F-tests are performed, followed by paired t-tests, where applicable (cp. overview Table XI on page 131). Then, value structure, i.e. scaled part-worth utilities and respective attribute importances are compared and commented on. When evaluating tests of hypotheses, we are looking for consistency of results over group comparison, model form, and performance measures, as well as on the magnitude of the effects. Having performed all tests, it was decided to present tables of value structure, i.e. part-worth utilities and attribute importances only for the best of the five individual-level model forms.

4.2.1 Reliability Over Time and Over Attribute Set

Research Question # 1.

> What is the influence of the type of attribute chosen for the evaluative task (i.e. technical or product-referent attributes and non-technical or user-referent attributes) on customer value structure and predictive validity ?

From the literature review, hypotheses for the attribute set may be stated as follows:

H_0: The inclusion of user-referent attributes in the attribute set does not increase predictive performance.

H_A: The inclusion of user-referent attributes in the attribute set does increase predictive performance.

In order to test this hypothesis, three different groups are compared and tested on their performance measures, one between-subjects comparison, and two within-subjects comparisons. Accordingly, three comparisons of value structure are presented.

Predictive Performance

Table XIV on page 141 gives an overview over different performance measures for between-subjects comparison of groups G5 and G6. Consistent across all performance measures, and for all model forms, G5 with attribute set A2 which comprises technical and one user-referent attribute set shows better performance than group G6 with the solely technical attribute set A1, suggesting increased predictive accuracy with the inclusion of user-referent attributes. In order to gauge believability of differences in performances, according F-tests for Fisher´s z, RMSE, and First-Hit are provided in Table XV on page 142.

Model fit R^2 for the calibration set of profiles ranges from 0.8693 to 0.8909 for G5, and from 0.8381 to 0.8766 for G6.. These differences cannot be tested across conjoint models and groups, as their magnitudes, significance, and power are design-dependent. With Umesh and Mishra´s (1990) Monté Carlo study, influences on the magnitude of R^2 using OLS regression are established for the number of profiles used for calibration, the number of attributes, and the distribution of importances among the attributes. Based on these selected influences, and Table 4´s entries (Umesh and Mishra 1990, p. 41) for thirty-two (32) profiles, eight (8) attributes, an equal weighted to moderately dominant importance distribution, and variances of about 25%, an R^2 between 0.864 and 1.000 may be termed excellent, and the range of 0.717 to 0.878 be called fair. A random model for these design parameters would receive an average R^2 of 0.568 at the 95% confidence level, and an R^2 of 0.518 at the 90% confidence level.

Power, i.e. the probability of rejecting a false null hypothesis, is over 99%. Therefore, one may be quite confident that, on average, these conjoint models provide for good to excellent models of customer value, esp. when considering DFs of 9 and 11 for models with and without interactions. However, no statistical inference about the relative superiority of different attribute sets, i.e. of the differences in R^2 , may be made, other than observing the consistent pattern of higher R^2 's for attribute set A2 including a non-technical, i.e. user-referent attribute.

Adj. R^2 also shows a consistent pattern of higher values for the mixed technical and user-referent attribute set A2. The differences are even more pronounced than for unadjusted R^2. In this case too, there is no valid statistical test for these differences.

Gauging predictive accuracy with R^2 on the holdout profiles, i.e. the variance accounted for with models based on attribute sets A1 and A2, predictive performance is markedly improved when including a user-referent attribute in the product description. First, absolute values of a low of 0.5211 to a high of 0.6003 for A2 over conjoint and self-explicated models are excellent in terms of variance explained, and considering the complexity of the conjoint task. Second, including a non-technical attribute in the profile description consistently explains from about ten (9.95) to fourteen (14.06) percentage points more variance than solely technical attribute set A1. As this result is based on the holdout profiles, and results are consistent across model forms, there is good evidence that inclusion of user-referent attributes increases predictive accuracy.

TABLE XIV

PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS; G5 / G6

Type of Model	Performance Measures (Averages Over Groups)								
	R^2 (calib.)	Adj R^2 (calib.)	R^2 (hold.)[b]	r_{xy} (hold.)[a]	Fisher's $z(r_{xy})$ (hold.)[a]	RMSE (hold.)	RMSE (calib.)	First-Hit (hold.)	First-Hit (mean counts)
1. TC main effects	0.8693 / 0.8381	0.6910 / 0.6173	0.6003 / 0.5008	0.7468 / 0.6925	1.0998 / 0.9121	19.35 / 21.72	7.24 / 8.52	70.83% / 65.00%	2.83 / 2.60
2. TC iAxD	0.8871 / 0.8766	0.6737 / 0.6435	0.5910 / 0.4692	0.7448 / 0.6598	1.0809 / 0.8584	19.69 / 22.66	6.75 / 7.47	73.33% / 65.83%	2.93 / 2.63
3. TC iBxD	0.8918 / 0.8706	0.6875 / 0.6261	0.5211 / 0.3974	0.6881 / 0.6027	0.9575 / 0.7441	22.21 / 27.11	6.37 / 7.54	69.17% / 60.00%	2.77 / 2.40
4. TC iCxD	0.8909 / 0.8628	0.6849 / 0.6035	0.5900 / 0.4866	0.7391 / 0.6792	1.0792 / 0.8842	19.80 / 21.92	6.62 / 7.88	70.83% / 64.17%	2.83 / 2.57
5. SE	n/a	n/a	0.5307 / 0.3901	0.7104 / 0.6013	0.9573 / 0.7428	26.14 / 28.61	n/a	70.83% / 57.50%	2.83 / 2.30

(calib.) = Calibration set
(hold.) = Holdout set

[a] = Seemingly non-monotone transformations between r_{xy} and $z(r_{xy})$ when comparing different cells in the table result from averaging individual results which is appropriate for Fisher's z, but not for r_{xy}.

[b] = Averaged from the individual R^2s (calculations from group r_{xy} show too low coefficients).

TABLE XV

F-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (G5 / G6 ; $F_{1,58}$ DEGREES OF FREEDOM)

Type of Model	F-Tests of Performance Measures (Averages Over Groups)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G5 / G6	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G5 / G6	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G5 / G6
1. TC main effects	3.5666 (.0640) 5.79%	[0.9597;1.2399] / [0.8179;1.0064]	1.4239 (.2376) 2.40%	[17.21;21.49] / [19.12;24.31]	0.9595 (.3314) 1.63%	[2.56;3.10] / [2.30;2.90]
2. TC iAxD	5.2075 (.0262) 8.24%	[0.9498;1.2120] / [0.7571;0.9597]	2.1129 (.1515) 3.51%	[17.58;21.81] / [19.91;25.41]	2.0163 (.1610) 3.36%	[2.68;3.19] / [2.38;2.88]
3. TC iBxD	4.8643 (.0314) 7.74%	[0.8195;1.0955] / [0.6549;0.8334]	3.4656 (.0677) 5.64%	[19.49;24.93] / [23.56;30.65]	2.3131 (.1337) 3.84%	[2.45;3.08] / [2.13;2.67]
4. TC iCxD	4.0002 (.0502) 6.45%	[0.9417;1.2168] / [0.7919;0.9766]	1.1330 (.2916) 1.92%	[17.63;21.98] / [19.34;24.50]	1.1557 (.2868) 1.95%	[2.55;3.12] / [2.25;2.88]
5. SE	7.1223 (.0099) 10.94%	[0.8568;1.0579] / [0.6503;0.8352]	0.5508 (.4610) 0.94%	[22.30;29.98] / [24.47;32.74]	5.1059 (.0276) 8.09%	[2.55;3.12] / [2.02;2.58]

(hold.) = Holdout set; [a] = Set of both groups; [b] = Group forms its own set (two-tailed, $\alpha = 0.1$; DFs 29/29)

Pearson product moment correlation r_{xy} between actual and predicted profile ratings suggest the same consistent pattern of improved prediction using attribute set A2 over A1 for different model forms. The differences are between about five and ten percentage points. However, one has to be cautious comparing these differences as the scale for the correlation coefficient r_{xy} is not interval-scaled, and differences at high values are actually larger than the same differences at low values. Unfortunately, there is no gauge to evaluate the magnitude of violation of interval scale for r_{xy}. Therefore, Fisher's z-transformation is applied in order to make the scale of r_{xy} (approximately) interval scaled (except for values at the extreme end of the scale), allowing for valid F-tests[20]. Absolute improvements are now more marked on the high end of the scale than on the low end. Table XV on page 142 provides results of F-tests on Fisher's z-transformed r_{xy}'s. The improvement in prediction from attribute set A1 to A2 is clearly significant for the SE model ($p < 0.0099$), significant at $p < 0.0262$ and $p < 0.0314$ for TC iAxD and TC iBxD, and only marginally (in)significant for the TC main effects and the TC iCxD conjoint models with $p < 0.0640$ and $p < 0.0502$, respectively. Providing separate intervals for 90% confidence into the mean group values shows wide margins for the ranges around the means, though with little overlap. In conclusion, one would hope for a clearer picture of the statistical tests for the increase in predictive accuracy provided with A2 that is demonstrated with the consistent picture of increased absolute values of predictive performance.

Absolute values of RMSE for both calibration and holdout set of profiles show the same consistent pattern of improvements in prediction with attribute set A2 over all

[20] Some authors that performed tests on both r_{xy} and Fisher's z report no change in substantive findings (cp. Green, Helsen, and Shandler 1988, footnote 4 on page 395). However, as this cannot be replicated due to missing data, and in order to avoid duplication of effort at later stages of this study, all ANOVAs are performed on Fisher's z-transformed values rather than on r_{xy}'s themselves.

model forms. Comparing absolute values of RMSE for calibration and holdout sets of profiles, the holdout sample shows about three (3) times the variation of the calibration set. Performing F-tests, the differences between the two sets of attributes are not significant to suggest belief these differences result from differences in the attribute sets. A possible resolution of the contradiction between consistent patterns of absolute measures and nonsignificant F-tests is provided later in this section.

Finally, evaluating first choice hit rates (First-Hit), again all five model forms with attribute set A2 show consistent improvement in predictive performance. In absolute terms, values around 70% correct predictions of first choice out of sets of four profiles per set may be considered very good in light of other conjoint and consumer research literature. However, comparing these figures to according F-tests on the mean counts of First-Hits, they show no significance for the improvement in prediction for the conjoint models, and a significance of $p < 0.0276$ for the improvement from the low value of 57.5% to 70.8% for the self-explicated (SE) model. However, checking for the reasons why this may be the case, it turns out that the distributional assumption of normality is violated for the First-Hit data across all model forms, including for the SE model[21]. Though F-tests are not very sensitive to violations of normality, the significant violation for First-Hit data of these groups may obscure small differences while still showing significance for large ones.

Summarizing results for between-subjects comparison and test for reliability over attribute set, all performance measures show a consistent pattern of improved reliability when including a user-referent attribute in the product description.

[21] Assumptions of normality of distribution of inputs for all F-tests were tested using the Shapiro-Wilk Wtest for normality. All p-values for the count data were less than 0.0023, leading to the conclusion that the distributions are not normal, and thus may be distorting the F-tests. Shapiro-Wilk's test is preferred over Kolmogorov-Smirnov, as it shows good power over a variety of situations (Norusis/SPSS 1993, p. 190)

However, this consistent picture is not mirrored in the F-tests on RMSEs and First-Hits, possibly because the differences are too small to be significant, i.e. the power of the tests may be too low for the effect to be detected. For First-Hits data, violation of the assumption of normal distribution of the inputs may be the cause for nonsignificance of the F-tests. However, for RMSEs this cannot be asserted, though there is also a tendency to deviate from normal distribution. Another explanation for a clearly distinguished pattern but marginal to absent significance in the F-tests may be that variations due to the attribute set are only a little smaller than individual respondent variation. In this case, the systematic influence of attribute set would show in the performance measures but may be obscured in the F-tests by the larger individual variation. In order to gauge individual variation, i.e. reliability over time which is sometimes termed the true error, paired t-tests were performed on the same groups G5 and G6, in addition to F-tests. Testing the difference between Part of G2, i.e. the first measurement of individuals in group G5, and G5 tests the difference that is solely due to time of administration. Testing the difference between Part of G2, i.e. the first measurement of individuals in group G6, and G6 tests the confounded difference due to time of administration and attribute set. Comparing results of these tests to the between-subjects F-tests provides some measure of the relative magnitudes of individual and treatment effects.

Table XVI on page 146 shows predictive performance of paired group G5.
Tables XVII and XVIII on pages 147 and 148 provide associated F-tests and paired t-tests, the latter of which are more appropriate for a repeated measurement. Here, improvements in prediction from the first to the second measurement are not as

TABLE XVI

PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS; PART G2 / G5

	Type of Model	Performance Measures (Averages Over Groups)								
		R^2 (calib.)	Adj R^2 (calib.)	R^2 (hold.) [b]	r_{xy} (hold.) [a]	Fisher's $z(r_{xy})$ (hold.) [a]	RMSE (hold.)	RMSE (calib.)	First-Hit (hold.)	First-Hit (mean counts)
1.	TC main effects	0.8559 / 0.8693	0.6594 / 0.6910	0.5614 / 0.6003	0.7325 / 0.7468	1.0689 / 1.0998	21.49 / 19.35	8.73 / 7.24	71.67% / 70.83%	2.87 / 2.83
2.	TC iAxD	0.8836 / 0.8871	0.6636 / 0.6737	0.5265 / 0.5910	0.7050 / 0.7448	1.0036 / 1.0809	22.24 / 19.69	7.89 / 6.75	68.33% / 73.33%	2.73 / 2.93
3.	TC iBxD	0.8814 / 0.8918	0.6574 / 0.6875	0.4465 / 0.5211	0.6428 / 0.6881	0.8415 / 0.9575	26.15 / 22.21	7.97 / 6.37	65.00% / 69.17%	2.60 / 2.77
4.	TC iCxD	0.8845 / 0.8909	0.6663 / 0.6849	0.5349 / 0.5900	0.7074 / 0.7391	1.0241 / 1.0792	22.07 / 19.80	7.88 / 6.62	70.83% / 70.83%	2.83 / 2.83
5.	SE	n/a	n/a	0.5321 / 0.5307	0.7106 / 0.7104	0.9711 / 0.9573	27.46 / 26.14	n/a	69.17% / 70.83%	2.77 / 2.83

(calib.) = Calibration set
(hold.) = Holdout set

[a] = Seemingly non-monotone transformations between r_{xy} and $z(r_{xy})$ when comparing different cells in the table result from averaging individual results which is appropriate for Fisher's z, but not for r_{xy}.

[b] = Averaged from the individual R^2s (calculations from group r_{xy} show too low coefficients).

TABLE XVII

F-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (PART G2 / G5 ; $F_{1,58}$ DEGREES OF FREEDOM)

Type of Model	F-Tests of Performance Measures (Averages Over Groups)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G2 / G5	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G2 / G5	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G2 / G5
1. TC main effects	0.0631 (.8026) 0.11%	[0.9144;1.2235] / [0.9597;1.2399]	1.0450 (.3109) 1.77%	[18.66;24.31] / [17.21;21.49]	0.0203 (.8872) 0.04%	[2.58;3.16] / [2.56;3.10]
2. TC iAxD	0.4301 (.5145) 0.74%	[0.8522;1.1550] / [0.9498;1.2120]	1.5150 (.2233) 2.55%	[19.43;25.06] / [17.58;21.81]	0.8339 (.3649) 1.42%	[2.46;3.00] / [2.68;3.19]
3. TC iBxD	1.1983 (.2782) 2.02%	[0.7258;0.9572] / [0.8195;1.0955]	2.7285 (.1040) 4.49%	[23.15;29.15] / [19.49;24.93]	0.4597 (.5004) 0.79%	[2.32;2.88] / [2.45;3.08]
4. TC iCxD	0.1987 (.6574) 0.34%	[0.8654;1.1829] / [0.9417;1.2168]	1.1148 (.2954) 1.89%	[19.15;24.98] / [17.63;21.98]	0.0000 (1.0000) 0.00%	[2.53;3.14] / [2.55;3.12]
5. SE	0.0242 (.8770) 0.04%	[0.8589;1.0834] / [0.8568;1.0579]	0.1768 (.6757) 0.30%	[23.79;31.13] / [22.30;29.98]	0.0781 (.7809) 0.13%	[2.48;3.06] / [2.55;3.12]

(hold.) = Holdout set; [a] = Set of both groups; [b] = Group forms its own set (two-tailed, $\alpha = 0.1$; DFs 29/29)

TABLE XVIII

T-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (PART G2 / G5 ; 29 DEGREES OF FREEDOM)

Type of Model	Paired t-tests of Performance Measures (Averages Over Groups)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
	Mean Diff. t-Ratio (Prob < t)	Conf. Int. (90%) [a] Diff. Part G2 / G5	Mean Diff. t-Ratio (Prob > t)	Conf. Int. (90%) [a] Diff. Part G2 / G5	Mean Diff. t-Ratio (Prob < t)	Conf. Int. (90%) [a] Diff. Part G2 / G5
1. TC main effects	-0.0308 -0.2593 (.3986)	[-0.1868; 0.1251]	2.1353 1.4423 (.0800)	[0.19; 4.08]	0.0333 0.1663 (.5655)	[-0.23; 0.30]
2. TC iAxD	-0.0773 -0.6732 (.2531)	[-0.2279; 0.0733]	2.5503 1.7112 (.0489)	[0.60; 4.50]	-0.2000 -1.0630 (.1483)	[-0.45; 0.05]
3. TC iBxD	-0.1160 -1.1516 (.1294)	[-0.2481; 0.0161]	3.9373 2.3091 (.0141)	[1.70; 6.17]	-0.1667 -0.7957 (.2163)	[-0.44; 0.11]
4. TC iCxD	-0.0551 -0.4576 (.3253)	[-0.2130; 0.1028]	2.2603 1.5049 (.0716)	[0.29; 4.23]	0.0000 0.0000 (.5000)	[-0.26; 0.26]
5. SE	0.0138 0.1942 (.5763)	[-0.0793; 0.1069]	1.3147 0.7122 (.2410)	[-1.11; 3.73]	-0.0667 -0.3283 (.3725)	[-0.33; 0.20]

(hold.) = Holdout set; Mean Diff. = (Average) difference between the two group(s) means

[a] = Set of group differences (one-tailed, $\alpha = 0.1$; DFs 29 ; $t^* = 1.311$)

consistent as with the improvement of the attribute set. Specifically, conjoint models and the SE model show a different pattern. For the four different conjoint models, 32 out of 36 comparisons over time show a slight improvement in prediction for the second measurement. Two comparisons show ties (First-Hit of TC iCxD) and two comparisons show a slight deterioration (First-Hit TC main effects). For the self-explicated model, the pattern is reversed: Out of six (6) measures, three (3) show a slight deterioration in predictive performance for the second measurement, whereas three (3) measures (RMSE and First-Hits) show slightly improved performance. All F-tests and all paired t-tests (except for two) for significance of observed improvements or deteriorations in prediction are not significant and relatively small in absolute values, strongly suggesting that conjoint analysis is reliable over time. Removing one extreme value from the second measurement of G5 also yields insignificance for the two significant paired t-tests. Furthermore, the differences between the two measurements over time are generally smaller than differences observed with changes in the attribute set, suggesting variations due to the attribute set are not smaller than those due to the individual respondents. This result, however, suggests that though there is some evidence for systematic improvement of prediction with the inclusion of user-referent attributes, the improvement is not large enough to *clearly* show in statistical tests.

Tables XIX, XX and XXI on pages 150 to 152 provide an overview over predictive performances of paired group G6 and associated F-tests and paired t-tests. The difference between these two measurements confounds effects of time with effects due to the attribute set. The second measurement is performed with attribute set A1, i.e. the solely technical attribute set which showed a consistent tendency of lower predictive accuracy than conjoint models with the mixed technical and user-referent

TABLE XIX

PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS; PART G2 / G6

	Type of Model	Performance Measures (Averages Over Groups)								
		R^2 (calib.)	Adj R^2 (calib.)	R^2 (hold.) [b]	r_{xy} (hold.) [a]	Fisher's $z(r_{xy})$ (hold.) [a]	RMSE (hold.)	RMSE (calib.)	First-Hit (hold.)	First-Hit (mean counts)
1.	TC main effects	0.8347 / 0.8381	0.6093 / 0.6173	0.4865 / 0.5008	0.6862 / 0.6925	0.8892 / 0.9121	23.99 / 21.72	8.53 / 8.52	71.67% / 65.00%	2.87 / 2.60
2.	TC iAxD	0.8647 / 0.8766	0.6091 / 0.6435	0.4629 / 0.4692	0.6596 / 0.6598	0.8607 / 0.8584	24.63 / 22.66	7.65 / 7.47	65.83% / 65.83%	2.63 / 2.63
3.	TC iBxD	0.8622 / 0.8706	0.6020 / 0.6261	0.4047 / 0.3974	0.6176 / 0.6027	0.7612 / 0.7441	28.80 / 27.11	7.73 / 7.54	64.17% / 60.00%	2.57 / 2.40
4.	TC iCxD	0.8602 / 0.8628	0.5963 / 0.6035	0.4617 / 0.4866	0.6648 / 0.6792	0.8516 / 0.8842	24.62 / 21.92	7.77 / 7.88	71.67% / 64.17%	2.87 / 2.57
5.	SE	n/a	n/a	0.4479 / 0.3901	0.6548 / 0.6013	0.8234 / 0.7428	26.46 / 28.61	n/a	64.17% / 57.50%	2.57 / 2.30

(calib.) = Calibration set
(hold.) = Holdout set

[a] = Seemingly non-monotone transformations between r_{xy} and $z(r_{xy})$ when comparing different cells in the table result from averaging individual results which is appropriate for Fisher's z, but not for r_{xy}.

[b] = Averaged from the individual R^2s (calculations from group r_{xy} show too low coefficients).

TABLE XX

F-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (PART G2 / G6 ; $F_{1,58}$ DEGREES OF FREEDOM)

	Type of Model	F-Tests of Performance Measures (Averages Over Groups)					
		Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
		$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G2 / G6	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G2 / G6	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G2 / G6
1.	TC main effects	0.0891 (.7663) 0.15%	[0.7993;0.9792] / [0.8179;1.0064]	1.1003 (.2985) 1.86%	[21.38;26.61] / [19.12;24.31]	1.3257 (.2543) 2.23%	[2.61;3.12] / [2.30;2.90]
2.	TC iAxD	0.0007 (.9792) 0.00%	[0.7528;0.9686] / [0.7571;0.9597]	0.7176 (.4004) 1.22%	[21.80;27.46] / [19.91;25.41]	0.0000 (1.0000) 0.00%	[2.31;2.95] / [2.38;2.88]
3.	TC iBxD	0.0570 (.8122) 0.10%	[0.6793;0.8430] / [0.6549;0.8334]	0.3476 (.5578) 0.60%	[25.44;32.16] / [23.56;30.65]	0.4976 (.4834) 0.85%	[2.27;2.87] / [2.13;2.67]
4.	TC iCxD	0.1822 (.6710) 0.31%	[0.7603;0.9430] / [0.7919;0.9766]	1.5324 (.2207) 2.57%	[21.96;27.28] / [19.34;24.50]	1.6034 (.2105) 2.69%	[2.61;3.12] / [2.25;2.88]
5.	SE	1.2681 (.2648) 2.14%	[0.7444;0.9024] / [0.6503;0.8352]	0.4880 (.4876) 0.83%	[23.28;29.64] / [24.47;32.74]	1.4168 (.2388) 2.38%	[2.31;2.82] / [2.02;2.58]

(hold.) = Holdout set; [a] = Set of both groups; [b] = Group forms its own set (two-tailed, $\alpha = 0.1$; DFs 29/29)

TABLE XXI

T-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (PART G2 / G6 ; 29 DEGREES OF FREEDOM)

		Paired t-tests of Performance Measures (Averages Over Groups)					
	Type of Model	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
		Mean Diff. t-Ratio (Prob < t)	Conf. Int. (90%) [a] Diff. Part G2 / G6	Mean Diff. t-Ratio (Prob > t)	Conf. Int. (90%) [a] Diff. Part G2 / G6	Mean Diff. t-Ratio (Prob < t)	Conf. Int. (90%) [a] Diff. Part G2 / G6
1.	TC main effects	-0.0229 -0.3666 (.7166)	[-0.1290; 0.0832]	2.2757 1.4528 (.1570)	[-0.39; 4.94]	0.2667 1.0922 (.2838)	[-0.15; 0.68]
2.	TC iAxD	0.0023 0.0328 (.9740)	[-0.1159; 0.1205]	1.9677 1.2503 (.2212)	[-0.71; 4.64]	0.0000 0.0000 (1.0000)	[-0.37; 0.37]
3.	TC iBxD	0.0170 0.2665 (.7917)	[-0.0914; 0.1255]	1.6943 0.7327 (.4696)	[-2.23; 5.62]	0.1667 0.7235 (.4752)	[-0.22; 0.56]
4.	TC iCxD	-0.0326 -0.5608 (.5792)	[-0.1315; 0.0662]	2.6997 1.6669 (.1063)	[-0.05; 5.45]	0.3000 1.2477 (.2221)	[-0.11; 0.71]
5.	SE	0.0806 1.3859 (.1763)	[-0.0182; 0.1794]	-2.1453 -1.2190 (.2327)	[-5.14; 0.84]	0.2667 1.2782 (.2113)	[-0.09; 0.62]

(hold.) = Holdout set; Mean Diff. = (Average) difference between the two group(s) means

[a] = Set of group differences (two-tailed, $\alpha = 0.1$; DFs 29 ; $t^* = 1.699$)

attribute set A2. However, a second measurement shows an even smaller tendency to increase accuracy in prediction.

Confounding these two effects should cancel them out. This is exactly what is visible in Table XIX on page 150. A little less than half of the 42 performance measures (19) show unchanged or better predictive performance for the first measurement with attribute set A2, while the other half (23) show better performance on the second measurement with attribute set A1: Effects of time and attribute set seem to cancel out. Formal tests show that all F-tests and paired t-tests on the differences for Fisher´s z, RMSE, and First-Hit are not significant (cp. Tables XX and XXI on pages 151 and 152). This, too, suggests strong evidence that conjoint measurement is reliable over attribute set.

In conclusion of results from the F-tests and the paired t-tests one cannot reject H_0 that the inclusion of user-referent attributes in the attribute set does not increase predictive performance at the $\alpha < 0.05$ level. As measures Fisher´s z (respectively correlations r_{xy}) and RMSE show significance at the $\alpha < 0.1$ level for the majority of them in the between-subjects tests, there is, however, a tendency for user-referent attributes to increase predictive accuracy. Nevertheless, conjoint analysis may safely be considered reliable over attribute set.

Value Structure

Value structure refers to part-worths and respective importances of attributes. In above tables showing performance measures, the conjoint model with main effects and without interaction terms showed overall best predictive performance over measures and measurement conditions. Therefore, value structure for all three comparisons concerning reliability over attribute set are presented for the main effects conjoint model, here.

Table XXII on page 155 shows scaled part-worths, importances, and importance ranks for the between-subjects comparison of groups G5 consisting of attribute set A2, and G6 consisting of attribute set A1. In agreement with consumer research literature, the three (3) most important attributes comprise over 50% of importance weights for all nine (9) attributes (54.7% for G5 and 50.0% for G6), i.e. the first three most important attributes explain over 50% of deviations in the response. Examining absolute importances, Base-Price is the most important attribute for attribute set A2 with 28.8%, as it is for attribute set A1 with 19.0%. Second in attribute importance is Features with 15.4% for G5 and 17.3% for G6., and third in importance is the user-referent attribute Firm-Reputation with 10.5% for G5, and the technical attribute Type-of-Display with 13.7% for G6. Considering differences in importances, the most marked effect is the significant deviation in the importance of price when the user-referent attribute is present ($F_{1,58} = 5.39$; $p < 0.0237$). Considering the eight other attributes, the difference between attribute importances is between one (1) and four (4) percentage points. This shows very high reliability over attributes for all seven technical attributes, and remarkably enough, only marginal deviations in importances for the perturbed attributes, too. However, some counter-intuitive deviations from expected level-utility functions occur with the inclusion of the user-referent attribute: For screen size and battery life in attribute set A2, the medium values of 9.4 inches and 5 hours are less preferred than the low values of 8.4 inches and 3 hours. In contrast to G5, these counter-intuitive attribute preferences do not occur for the solely technical attribute set A1: There, level-utility functions are in accordance with expectations. Finally, the user-referent attribute Firm-Reputation, as expected, shows a monotone level-utility function. The level utilities for the attribute PointDev

TABLE XXII

PART-WORTHS AND ATTRIBUTE IMPORTANCES G5 / G6 (TC MAIN EFFECTS)

Attribute Levels: Coded (Actual)	Scaled Part-Worths		Importance		Importance Rank
	G5	G6	G5	G6	G5 / G6
Weight-1 (9 pounds [a])	-0.3505	-0.6656			
Weight0 (7 pounds)	0.2902	0.2161			
Weight1 (5 pounds [a])	0.0604	0.4494	7.34%	8.54%	7 / 6
ScrSiz-1 (8.4 inch diagonal [a])	-0.0154	-0.1153			
ScrSiz0 (9.4 inch diagonal)	-0.2036	-0.1165			
ScrSiz1 (10.4 inch diagonal[a])	0.2190	0.2319	7.47%	7.54%	6 / 8
Display-1 (Monochrome)	-1.1342	-1.5987			
Display1 (Color)	1.1342	1.5987	10.22%	13.74%	4 / 3
B_Price-1 ($ 3500 [a])	-3.4202	-2.1189			
B_Price0 ($ 2500)	0.2849	0.0936			
B_Price1 ($ 1500 [a])	3.1353	2.0253	28.84%	19.01%	1 / 1
Keyb_Siz-1 (Smaller than regular size)	-0.0211	-0.1483			
Keyb_Siz1 (Regular size)	0.0211	0.1483	4.37%	5.93%	9 / 9
BattLife-1 (3 hours[a])	-0.0962	-0.5971			
BattLife0 (5 hours)	-0.2178	0.0357			
BattLife1 (7 hours[a])	0.3140	0.5614	9.17%	10.88%	5 / 4
Speed-1 (Comfortable for word-processing)	-0.1576	-0.3758			
Speed1 (Fast for big spreadsheet and imaging)	0.1576	0.3758	6.71%	7.83%	8 / 7
Features-1 (No additional features[a])	-1.6968	-1.9257			
Features0 (Expansion slots for keyboard, monitor, others)	0.5716	0.1082			
Features1 (Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others[a])	1.1252	1.8176	15.37%	17.29%	2 / 2
Firm_Rep-1 / PointDev-1	-0.7544	0.3466			
Firm_Rep0 / PointDev0	-0.0278	-0.2181			
Firm_Rep1 / PointDev1	0.7822	-0.1285	10.52%	9.25%	3 / 5

[a] Levels used for the 2-level extreme design of the holdout product profiles.

(pointing device) cannot be assumed to be monotone. Therefore, these part-worth utilities do not contradict the statements just made above.

In order to gauge if value structure for between-subjects effects of attribute sets A2 and A1 are not caused by subject variation, within-subject comparisons of value structure are compiled in Table XXIII on page 157 for paired first and second measurements of individuals in group G5, i.e. with attribute set A2 resulting in some counter-intuitive level utilities, and in Table XXIV on page 159 for paired group G6. The paired comparison for group G5 contrasts differences solely due to time of administration, whereas the paired comparison of G6 shows differences due to time confounded with differences due to change from attribute set A2 in the first measurement to attribute set A1 in the second measurement.

Importances of the first measurement for group G5 with attribute set A2 are very close to those of the second measurement. Again, the three most important attributes of the first measurement — Base-Price, Features, and Type-of-Display, respectively — comprise more than half of the importance weights (52.6%). In terms of absolute differences in importance for all nine attributes, no difference is greater than Base-Price´s 3.68 percentage points with most of the rest below the one (1) percentage point mark. Additionally, changes of importance ranks occur only for four attributes, and then only for adjacent ranks. Both of the latter observations strongly suggest high reliability of value structure over time of administration. A possibly problematic outcome of the first measurement, however, is the fact that the counter-intuitive level-utility functions of the second measurement are not present in the first. Nevertheless, as these deviations in level-utility functions only occur in the least important ones and they do not reverse the order of best and worst level utility for the monotone and

TABLE XXIII

PART-WORTHS AND ATTRIBUTE IMPORTANCES PART G2 / G5 (TC MAIN EFFECTS)

Attribute Levels: Coded (Actual)	Scaled Part-Worths		Importance		Importance Rank
	Part G2	G5	Part G2	G5	Part G2/G5
Weight-1 (9 pounds [a])	-0.1772	-0.3505			
Weight0 (7 pounds)	-0.0229	0.2902			
Weight1 (5 pounds [a])	0.2001	0.0604	7.63%	7.34%	6 / 7
ScrSiz-1 (8.4 inch diagonal [a])	-0.2275	-0.0154			
ScrSiz0 (9.4 inch diagonal)	-0.0549	-0.2036			
ScrSiz1 (10.4 inch diagonal[a])	0.2824	0.2190	7.04%	7.47%	7 / 6
Display-1 (Monochrome)	-1.2660	-1.1342			
Display1 (Color)	1.2660	1.1342	12.06%	10.22%	3 / 4
B_Price-1 ($ 3500 [a])	-2.8539	-3.4202			
B_Price0 ($ 2500)	0.2082	0.2849			
B_Price1 ($ 1500 [a])	2.6456	3.1353	25.16%	28.84%	1 / 1
Keyb_Siz-1 (Smaller than regular size)	0.0444	-0.0211			
Keyb_Siz1 (Regular size)	-0.0444	0.0211	5.15%	4.37%	9 / 9
BattLife-1 (3 hours[a])	-0.2489	-0.0962			
BattLife0 (5 hours)	-0.1052	-0.2178			
BattLife1 (7 hours[a])	0.3541	0.3140	9.14%	9.17%	5 / 5
Speed-1 (Comfortable for word-processing)	-0.1234	-0.1576			
Speed1 (Fast for big spreadsheet and imaging)	0.1234	0.1576	6.41%	6.71%	8 / 8
Features-1 (No additional features[a])	-1.9839	-1.6968			
Features0 (Expansion slots for keyboard, monitor, others)	0.6352	0.5716			
Features1 (Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others[a])	1.3487	1.1252	15.43%	15.37%	2 / 2
Firm_Rep-1 (No-name [a])	-0.8815	-0.7544			
Firm_Rep0 (Store brand)	-0.2783	-0.0278			
Firm_Rep1 (Well-known brand[a])	1.1598	0.7822	11.98%	10.52%	4 / 3

[a] Levels used for the 2-level extreme design of the holdout product profiles.

ordinal attributes (cp. Table X on page 99), this may not be considered as a problem for reliability over attribute set in terms of value structure.

Comparing the differences in importance between first and second administration for group G6 in Table XXIV on page 159, the consistent picture of differences solely due to time is only slightly disturbed. Differences of little more than four (4) percentage points are observed only for Features and the perturbation between attributes Firm-Reputation and Pointing-Device. Another remarkable difference is the relatively low importance of Base-Price for the first measurement of group G6 which is not in accordance with the higher value in both measurements for G5. As for the first measurement of G5 with an importance in Base-Price of 25.16%, the difference to the first measurement of group G6 of 20.87% cannot be attributed to the time of administration or to the attribute set (A2) which both were the same for both of these groups. The difference between both values of Base-Price for both groups´ first measurement (25.16% vs. 20.87%), however, is not significant ($F_{1,58} = 1.38$; $p < 0.2448$), suggesting that the low importance for Base-Price in the first measurement of G6 is a random effect. The rest of the attributes, however, do not show more than two (2) percentage points deviation between importances. Importance rank deviations, though, are slightly more characteristic than in the former two cases. However, concerning all these results, value structure may also be considered reliable over attribute set for the confounded effect.

TABLE XXIV

PART-WORTHS AND ATTRIBUTE IMPORTANCES PART G2 / G6 (TC MAIN EFFECTS)

Attribute Levels: Coded (Actual)	Scaled Part-Worths		Importance		Importance Rank
	Part G2	G6	Part G2	G6	Part G2/G6
Weight-1 (9 pounds [a])	-0.5803	-0.6656			
Weight0 (7 pounds)	0.1235	0.2161			
Weight1 (5 pounds [a])	0.4568	0.4494	9.27%	8.54%	5 / 6
ScrSiz-1 (8.4 inch diagonal [a])	-0.2440	-0.1153			
ScrSiz0 (9.4 inch diagonal)	-0.1022	-0.1165			
ScrSiz1 (10.4 inch diagonal[a])	0.3462	0.2319	7.78%	7.54%	7 / 8
Display-1 (Monochrome)	-1.6007	-1.5987			
Display1 (Color)	1.6007	1.5987	13.70%	13.74%	2 / 3
B_Price-1 ($ 3500 [a])	-2.2209	-2.1189			
B_Price0 ($ 2500)	-0.1656	0.0936			
B_Price1 ($ 1500 [a])	2.3865	2.0253	20.87%	19.01%	1 / 1
Keyb_Siz-1 (Smaller than regular size)	0.0193	-0.1483			
Keyb_Siz1 (Regular size)	-0.0193	0.1483	4.95%	5.93%	9 / 9
BattLife-1 (3 hours[a])	-0.4274	-0.5971			
BattLife0 (5 hours)	0.0728	0.0357			
BattLife1 (7 hours[a])	0.3546	0.5614	8.88%	10.88%	6 / 4
Speed-1 (Comfortable for word-processing)	-0.2111	-0.3758			
Speed1 (Fast for big spreadsheet and imaging)	0.2111	0.3758	7.66%	7.83%	8 / 7
Features-1 (No additional features[a])	-1.5912	-1.9257			
Features0 (Expansion slots for keyboard, monitor, others)	0.2348	0.1082			
Features1 (Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others[a])	1.3564	1.8176	13.22%	17.29%	4 / 2
Firm_Rep-1 / PointDev-1	-1.2037	0.3466			
Firm_Rep0 / PointDev0	0.0957	-0.2181			
Firm_Rep1 / PointDev1	1.1080	-0.1285	13.68%	9.25%	3 / 5

[a] Levels used for the 2-level extreme design of the holdout product profiles.

Finally, another important observation concerning value structure may be made with importances of SE models in Table XIII on page 137 for comparisons made above. For all groups, i.e. treatments, the SE models tend to produce average importances which are close to a random model. Specifically, SE models of groups G5 and G6 fail to recognize the shift in importance in price with the inclusion of the user-referent attribute into the profile description. Also, they do not show shifts in importances of attributes Features and the perturbed attributes. As it is very unlikely that all attributes are about equal in importance over all treatment groups, and as such a situation is not distinguishable from a random model, no inference about reliability over attribute set with respect to SE value structure may be made. However, one may make inferences, or at least speculate about the relative superiority of individual-level models. TC individual-level models, it seems, are more able to detect and gauge shifts in value structure and importances of attributes, than are SE models.

4.2.2 Reliability Over Time and Over Stimulus Set

Research Question # 2.

What is the influence of specific factorial designs, i.e. of specific combinations of product attribute values, on estimation of customer value structure and predictive accuracy ?

From the literature review, hypotheses for the stimulus set may be stated as follows:

H_0: The utilization of a specific fractional factorial design does not influence predictive performance.

H_A: The utilization of a specific fractional factorial design does influence predictive performance.

In order to test this hypothesis, once again, three different groups are compared and tested on their performance measures, one between-subjects comparison, and two within-subjects comparisons. Accordingly, three comparisons of value structure are presented.

Predictive Performance

Table XXV on page 162 gives an overview over different performance measures for between-subjects comparison of groups G3 and G4. This comparison gauges differences due to the different fractional factorial designs (FF1 and FF2), i.e. due to different stimulus sets. Consistent across all performance measures, and for all model forms, performance on the calibration set, i.e. model fit, for the first factorial design in group G3 is better than for the second factorial design of group G4 (cp. R^2 calib., Adj R^2 calib., and RMSE calib.). This suggests a slightly more efficient or more balanced design of FF1 than of FF2. Both designs as well as their derivations are provided in Appendix IV. When examining measures for the holdout set of profiles, the results are mixed with only RMSE showing deterioration in predictive accuracy for FF2 over all model forms, though this is also the case for some of the other performance measures on the holdouts. In order to gauge believability of differences in performance, according F-tests for Fisher´s z, RMSE, and First-Hit are provided in Table XXVI on page 163. All test results show no significance for differences in predictive accuracy between groups G3 and G4 with fractional factorials FF1 and FF2, respectively. These results suggest good reliability over stimulus set for conjoint models.

Evaluating absolute magnitudes of performance measures, both groups show results similar to group G5 which used the different set of attributes A2, and generally lower results for G6 which used the same set of attributes A1. The latter result may be due to the different sets of attributes in the first measurement. In general, however,

TABLE XXV

PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS; G3 / G4

	Type of Model	Performance Measures (Averages Over Groups)								
		R^2 (calib.)	Adj R^2 (calib.)	R^2 (hold.) [b]	r_{xy} (hold.) [a]	Fisher's $z(r_{xy})$ (hold.) [a]	RMSE (hold.)	RMSE (calib.)	First-Hit (hold.)	First-Hit (mean counts)
1.	TC main effects	0.8730 / 0.8548	0.6999 / 0.6569	0.5905 / 0.5963	0.7466 / 0.7615	1.0895 / 1.0672	17.34 / 20.11	7.16 / 8.54	72.22% / 74.17%	2.89 / 2.97
2.	TC iAxD	0.8912 / 0.8778	0.6857 / 0.6470	0.5602 / 0.5650	0.7256 / 0.7379	1.0324 / 1.0179	17.92 / 21.00	6.61 / 7.80	71.30% / 71.67%	2.85 / 2.87
3.	TC iBxD	0.8984 / 0.8767	0.7065 / 0.6438	0.4616 / 0.5013	0.6418 / 0.6896	0.8666 / 0.9182	22.99 / 23.40	6.36 / 7.83	58.33% / 61.67%	2.33 / 2.47
4.	TC iCxD	0.8915 / 0.8821	0.6865 / 0.6593	0.5691 / 0.5478	0.7328 / 0.7236	1.0516 / 0.9960	17.79 / 21.59	6.63 / 7.56	72.22% / 70.00%	2.89 / 2.80
5.	SE	n/a	n/a	0.4517 / 0.4825	0.6551 / 0.6565	0.8315 / 0.8794	26.25 / 29.00	n/a	66.67% / 67.50%	2.67 / 2.70

(calib.) = Calibration set
(hold.) = Holdout set

[a] = Seemingly non-monotone transformations between r_{xy} and $z(r_{xy})$ when comparing different cells in the table result from averaging individual results which is appropriate for Fisher's z, but not for r_{xy} .

[b] = Averaged from the individual R^2s (calculations from group r_{xy} show too low coefficients).

TABLE XXVI

F-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (G3 / G4 ; $F_{1,55}$ DEGREES OF FREEDOM)

Type of Model	F-Tests of Performance Measures (Averages Over Groups)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
	$F_{1,55}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G3 / G4	$F_{1,55}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G3 / G4	$F_{1,55}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G3 / G4
1. TC main effects	0.0503 (.8234) 0.09%	[0.9445;1.2346] / [0.9731;1.1613]	2.3106 (.1342) 4.03%	[14.95;19.73] / [18.10;22.13]	0.1327 (.7171) 0.24%	[2.63;3.15] / [2.72;3.22]
2. TC iAxD	0.0210 (.8854) 0.04%	[0.8920;1.1727] / [0.9184;1.1175]	3.0414 (.0868) 5.24%	[15.55;20.29] / [19.10;22.91]	0.0032 (.9548) 0.01%	[2.53;3.18] / [2.56;3.17]
3. TC iBxD	0.2530 (.6170) 0.46%	[0.7230;1.0102] / [0.8142;1.0222]	0.0299 (.8635) 0.05%	[19.46;26.51] / [21.13;25.68]	0.2599 (.6122) 0.47%	[1.96;2.71] / [2.21;2.72]
4. TC iCxD	0.2899 (.5925) 0.52%	[0.9091;1.1941] / [0.8891;1.1028]	3.9538 (.0517) 6.71%	[15.41;20.18] / [19.37;23.81]	0.1741 (.6781) 0.32%	[2.64;3.14] / [2.54;3.06]
5. SE	0.2736 (.6030) 0.50%	[0.7393;0.9238] / [0.7572;1.0017]	1.1187 (.2948) 1.99%	[22.70;29.79] / [26.25;31.75]	0.0124 (.9119) 0.02%	[2.35;2.98] / [2.31;3.09]

(hold.) = Holdout set; [a] = Set of both groups; [b] = Group forms its own set (two-tailed. $\alpha = 0.1$; DFs 26/29) absolute magnitudes of measures may be termed very good. Model fit R^2 for the calibration set of profiles, for instance, ranges from 0.8730 to 0.8984 for G3, and from 0.8548 to 0.8821 for G4. This may be termed excellent with respect to this conjoint study's design parameters, and when gauged with Umesh and Mishra's (1990) Monté Carlo study. The portion of variance accounted for with the model, and evaluated with the holdout set of profiles, reaches nearly 60% for the best model (59.6% for TC main effects), and the percentage of correctly predicted first choices reaches over 70% for most models (best with TC main effects of 74.2% correctly predicted choices in group G4).

In order to evaluate if high individual variation may have cancelled out systematic effects due to the fractional factorials FF1 and FF2, paired comparisons between the first and second measurements of groups G3 and G4 are performed and respective performance measures on predictive accuracy are contrasted in Table XXVII on page 165 for group G3, and in Table XXX on page 170 for group G4. Accordingly, F-tests and paired t-tests are provided for both groups. Tests in Tables XXVIII on page 166 and XXIX on page 167 for group G3 gauge significance of differences solely due to time of administration. Tests in Tables XXXI on page 171 and XXXII on page 172 for group G4 are performed on differences due to the confounded effects of time of administration and the change from factorial set FF1 to FF2 in the second measurement.

When comparing predictive performance and associated tests for reliability over time of group G3, a pattern emerges, differentiated along the level of accuracy in prediction reached, and the form of preference model employed. Consistent for all 42 measures for prediction, the second measurement yielded higher accuracy in prediction than the

TABLE XXVII

PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS; PART G1 / G3

	Type of Model	Performance Measures (Averages Over Groups)								
		R^2 (calib.)	Adj R^2 (calib.)	R^2 (hold.)[b]	r_{xy} (hold.)[a]	Fisher's $z(r_{xy})$ (hold.)[a]	RMSE (hold.)	RMSE (calib.)	First-Hit (hold.)	First-Hit (mean counts)
1.	TC main effects	0.8181 / 0.8730	0.5700 / 0.6999	0.4237 / 0.5905	0.6149 / 0.7466	0.8059 / 1.0895	22.96 / 17.34	9.31 / 7.16	62.04% / 72.22%	2.48 / 2.89
2.	TC iAxD	0.8533 / 0.8912	0.5762 / 0.6857	0.4153 / 0.5602	0.6069 / 0.7256	0.7897 / 1.0324	23.80 / 17.92	8.37 / 6.61	61.11% / 71.30%	2.44 / 2.85
3.	TC iBxD	0.8494 / 0.8984	0.5650 / 0.7065	0.3668 / 0.4616	0.5799 / 0.6418	0.7099 / 0.8666	26.00 / 22.99	8.46 / 6.36	56.48% / 58.33%	2.26 / 2.33
4.	TC iCxD	0.8429 / 0.8915	0.5462 / 0.6865	0.4207 / 0.5691	0.6110 / 0.7328	0.8110 / 1.0516	23.26 / 17.79	8.62 / 6.63	62.96% / 72.22%	2.52 / 2.89
5.	SE	n/a	n/a	0.3896 / 0.4517	0.5910 / 0.6551	0.7349 / 0.8315	27.68 / 26.25	n/a	63.89% / 66.67%	2.56 / 2.67

(calib.) = Calibration set
(hold.) = Holdout set

[a] = Seemingly non-monotone transformations between r_{xy} and $z(r_{xy})$ when comparing different cells in the table result from averaging individual results which is appropriate for Fisher's z, but not for r_{xy}.

[b] = Averaged from the individual R^2s (calculations from group r_{xy} show too low coefficients).

TABLE XXVIII

F-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (PART G1 / G3 ; $F_{1,52}$ DEGREES OF FREEDOM)

Type of Model	F-Tests of Performance Measures (Averages Over Groups)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
	$F_{1,52}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G1 / G3	$F_{1,52}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G1 / G3	$F_{1,52}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G1 / G3
1. TC main effects	6.1055 (.0168) 10.51%	[0.6743;0.9375] / [0.9445;1.2346]	7.5395 (.0083) 12.66%	[20.42;25.50] / [14.95;19.73]	2.8140 (.0994) 5.13%	[2.16;2.80] / [2.63;3.15]
2. TC iAxD	4.6753 (.0352) 8.25%	[0.6595;0.9199] / [0.8920;1.1727]	8.3998 (.0055) 13.91%	[21.27;26.32] / [15.55;20.29]	2.4237 (.1256) 4.45%	[2.14;2.75] / [2.53;3.18]
3. TC iBxD	2.3793 (.1290) 4.38%	[0.6130;0.8069] / [0.7230;1.0102]	1.3591 (.2490) 2.55%	[23.35;28.65] / [19.46;26.51]	0.0630 (.8029) 0.12%	[1.92;2.60] / [1.96;2.71]
4. TC iCxD	4.1777 (.0460) 7.44%	[0.6695;0.9525] / [0.9091;1.1941]	6.8469 (.0116) 11.64%	[20.61;25.92] / [15.41;20.18]	2.3256 (.1333) 4.28%	[2.19;2.85] / [2.64;3.14]
5. SE	1.3889 (.2440) 2.60%	[0.6297;0.8401] / [0.7393;0.9238]	0.2201 (.6409) 0.42%	[23.85;31.51] / [22.70;29.79]	0.1940 (.6614) 0.37%	[2.26;2.85] / [2.35;2.98]

(hold.) = Holdout set; [a] = Set of both groups; [b] = Group forms its own set (two-tailed, $\alpha = 0.1$; DFs 26/26)

TABLE XXIX

T-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (PART G1 / G3 ; 26 DEGREES OF FREEDOM)

Type of Model	Paired t-tests of Performance Measures (Averages Over Groups)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
	Mean Diff. t-Ratio (Prob < t)	Conf. Int. (90%) [a] Diff. Part G1 / G3	Mean Diff. t-Ratio (Prob > t)	Conf. Int. (90%) [a] Diff. Part G1 / G3	Mean Diff. t-Ratio (Prob < t)	Conf. Int. (90%) [a] Diff. Part G1 / G3
1. TC main effects	-0.2837 -2.9141 (.0036)	[-0.4117; -0.1556]	5.6196 3.6249 (.0006)	[3.58; 7.66]	-0.4074 -1.8373 (.0388)	[-0.70; -0.12]
2. TC iAxD	-0.2426 -2.5034 (.0095)	[-0.3700; -0.1152]	5.8789 3.8920 (.0003)	[3.89; 7.87]	-0.4074 -1.6958 (.0509)	[-0.72; -0.09]
3. TC iBxD	-0.1567 -1.9582 (.0305)	[-0.2618; -0.0515]	3.0152 1.6725 (.0532)	[0.64; 5.39]	-0.0741 -0.3373 (.3693)	[-0.36; 0.21]
4. TC iCxD	-0.2406 -2.3543 (.0132)	[-0.3750; -0.1062]	5.4722 3.4458 (.0010)	[3.38; 7.56]	-0.3704 -1.7268 (.0480)	[-0.65; -0.09]
5. SE	-0.0967 -1.4255 (.0830)	[-0.1858; -0.0075]	1.4356 0.6790 (.2516)	[-1.34; 4.22]	-0.1111 -0.4616 (.3241)	[-0.43; 0.21]

(hold.) = Holdout set; Mean Diff. = (Average) difference between the two group(s) means

[a] = Set of group differences (one-tailed, $\alpha = 0.1$; DFs 26 ; $t^* = 1.315$)

first measurement (cp. Table XXVII on page 165), which may be expected due to increased task familiarity. Accordingly, over model forms, the worst performing models (TC iBxD and SE) yielded no significant improvements from first to second measurements for the F-tests, and only one marginally significant improvement for the paired t-tests ($p < 0.0305$ for TC iBxD) which provide the stronger tests. However, the other three model forms yielded significant improvements in accuracy of prediction for the second measurement on Fisher´s z and RMSE. Distributional assumptions are not violated, i.e. they cannot be responsible for significance. Also, performing sensitivity analyses where the lowest response in the first measurement and the two highest responses in the second measurement are eliminated from the analysis, still yields marginal significance for the stronger paired t-tests on Fisher´s z and RMSE for the three best models. When further examining the absolute magnitudes of performance measures for the second response, they are in line with responses of other groups, i.e. a little worse than G5 with attribute set A2, and better than G4 and G6. However, when examining the first measurement, the percentage of variance explained is only 42.37%, improving to 59.05% in the second measurement. Compared to the other group responses, the first measurement is very low at about five (5) percentage points lower than the first measurement of G4 (cp. Table XXX on page 170). This may reflect some unfortunate random influences for this group in the first measurement. Together with very good responses in the second measurement this may have caused the significant improvement in predictive accuracy. In conclusion, reliability over time may be dependent on the level of accuracy in prediction already reached, with the potential of significant improvements with a second measurement when levels in terms of Fisher´s z (i.e. correlations) and RMSE are low, and the respondent task is difficult.

Comparing predictive performance and associated tests for confounded effects of time and stimulus set of group G4 in Table XXX on page 170, a pattern for different performance measures emerges. R^2 and Adj R^2 for the calibration set both show slight deterioration in performance, i.e. model fit, over all conjoint models for FF2, possibly reflecting the overall better efficiency or balance of fractional factorial FF1 over FF2. Nevertheless, in all cases, this deterioration is less than one (1) percentage point at an overall fit of over 85%, suggesting no meaningful effect. However, all other measures except for one (RMSE of the SE model for the holdout set of profiles), show an improvement of the second measurement over the first. Additionally, these improvements show a similar pattern of significance as the one for paired group G3, though not as pronounced. Fisher´s z and RMSE are marginally (in)significant for the F-test on models with high predictive accuracy, and more markedly significant for the stronger paired t-test for the same group of measures and model forms (TC main effects, TC iAxD, and TC iBxD). In the paired t-test the improvement from first to second measurement is also significant for the best two First-Hit measures. These results suggest that possible deterioration in prediction for fractional factorial FF2 has nearly no effect on the improvement of prediction with the second measurement.

In sum, considering the between-subjects and within-subjects comparisons, though different fractional factorials may show some systematic effect, it is not large enough to be significant, and definitely smaller or not recognizable when compared to deviations in measurements over time. Thus, one cannot reject H_0 that the utilization of a specific fractional factorial design does not influence predictive performance. It is hard to detect any effect from different orthogonal fractional factorials, at all. Conjoint analysis may safely be regarded as reliable over stimulus set, i.e. for different fractional factorial designs.

TABLE XXX

PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS; PART G1 / G4

Type of Model	Performance Measures (Averages Over Groups)								
	R^2 (calib.)	Adj R^2 (calib.)	R^2 (hold.)[b]	r_{xy} (hold.)[a]	Fisher's $z(r_{xy})$ (hold.)[a]	RMSE (hold.)	RMSE (calib.)	First-Hit (hold.)	First-Hit (mean counts)
1. TC main effects	0.8603 / 0.8548	0.6698 / 0.6569	0.4757 / 0.5963	0.6600 / 0.7615	0.8963 / 1.0672	24.60 / 20.11	9.33 / 8.54	65.00% / 74.17%	2.60 / 2.97
2. TC iAxD	0.8793 / 0.8778	0.6512 / 0.6470	0.4466 / 0.5650	0.6381 / 0.7379	0.8289 / 1.0179	25.54 / 21.00	8.67 / 7.80	60.00% / 71.67%	2.40 / 2.87
3. TC iBxD	0.8832 / 0.8767	0.6626 / 0.6438	0.3738 / 0.5013	0.5775 / 0.6896	0.7321 / 0.9182	30.23 / 23.40	8.49 / 7.83	57.50% / 61.67%	2.30 / 2.47
4. TC iCxD	0.8846 / 0.8821	0.6667 / 0.6593	0.4925 / 0.5478	0.6750 / 0.7236	0.9299 / 0.9960	24.51 / 21.59	8.43 / 7.56	66.67% / 70.00%	2.67 / 2.80
5. SE	n/a	n/a	0.4093 / 0.4825	0.5960 / 0.6565	0.7678 / 0.8794	28.95 / 29.00	n/a	62.50% / 67.50%	2.50 / 2.70

(calib.) = Calibration set
(hold.) = Holdout set

[a] = Seemingly non-monotone transformations between r_{xy} and $z(r_{xy})$ when comparing different cells in the table result from averaging individual results which is appropriate for Fisher's z, but not for r_{xy}.

[b] = Averaged from the individual R^2s (calculations from group r_{xy} show too low coefficients).

TABLE XXXI

F-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (PART G1 / G4 ; $F_{1,58}$ DEGREES OF FREEDOM)

	Type of Model	F-Tests of Performance Measures (Averages Over Groups)					
		Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
		$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G1 / G4	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G1 / G4	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] Part G1 / G4
1.	TC main effects	3.1088 (.0831) 5.09%	[0.7611;1.0314] / [0.9731;1.1613]	4.8165 (.0322) 7.67%	[21.77;27.43] / [18.10;22.13]	2.5336 (.1169) 4.19%	[2.30;2.90] / [2.72;3.22]
2.	TC iAxD	4.6139 (.0359) 7.37%	[0.7174;0.9404] / [0.9184;1.1175]	5.2253 (.0259) 8.26%	[22.76;28.33] / [19.10;22.91]	2.9299 (.0923) 4.81%	[2.05;2.75] / [2.56;3.17]
3.	TC iBxD	3.9834 (.0507) 6.43%	[0.6125;0.8516] / [0.8142;1.0222]	9.3060 (.0034) 13.83%	[27.18;33.28] / [21.13;25.68]	0.4856 (.4887) 0.83%	[1.98;2.62] / [2.21;2.72]
4.	TC iCxD	0.4106 (.5242) 0.70%	[0.7909;1.0688] / [0.8891;1.1028]	1.9027 (.1731) 3.18%	[21.69;27.34] / [19.37;23.81]	0.3258 (.5703) 0.56%	[2.37;2.96] / [2.54;3.06]
5.	SE	1.2374 (.2706) 2.09%	[0.6488;0.8867] / [0.7572;1.0017]	0.0003 (.9856) 0.00%	[25.65;32.26] / [26.25;31.75]	0.3633 (.5491) 0.62%	[2.09;2.91] / [2.31;3.09]

(hold.) = Holdout set; [a] = Set of both groups; [b] = Group forms its own set (two-tailed, $\alpha = 0.1$; DFs 29/29)

TABLE XXXII

T-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (PART G1 / G4 ; 29 DEGREES OF FREEDOM)

Type of Model	Paired t-tests of Performance Measures (Averages Over Groups)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
	Mean Diff. t-Ratio (Prob < t)	Conf. Int. (90%) [a] Diff. Part G1 / G4	Mean Diff. t-Ratio (Prob > t)	Conf. Int. (90%) [a] Diff. Part G1 / G4	Mean Diff. t-Ratio (Prob < t)	Conf. Int. (90%) [a] Diff. Part G1 / G4
1. TC main effects	-0.1709 -2.4913 (.0093)	[-0.2609; -0.0810]	4.4870 2.7476 (.0051)	[2.35; 6.63]	-0.3667 -1.8836 (.0348)	[-0.62; -0.11]
2. TC iAxD	-0.1890 -3.5820 (.0006)	[-0.2582; -0.1198]	4.5407 3.2302 (.0015)	[2.70; 6.38]	-0.4667 -1.9571 (.0300)	[-0.78; -0.15]
3. TC iBxD	-0.1861 -3.3428 (.0011)	[-0.2591; -0.1131]	6.8280 4.0121 (.0002)	[4.60; 9.06]	-0.1667 -0.7571 (.2276)	[-0.46; 0.12]
4. TC iCxD	-0.0661 -1.0199 (.1581)	[-0.1511; 0.0189]	2.9177 1.8823 (.0349)	[0.89; 4.95]	-0.1333 -0.6425 (.2628)	[-0.41; 0.14]
5. SE	-0.1117 -1.3873 (.0880)	[-0.2172; -0.0061]	-0.0460 -0.0280 (.5111)	[-2.20; 2.11]	-0.2000 -0.7693 (.2240)	[-0.54; 0.14]

(hold.) = Holdout set; Mean Diff. = (Average) difference between the two group(s) means

[a] = Set of group differences (one-tailed, $\alpha = 0.1$; DFs 29 ; $t^* = 1.311$)

Value Structure

As in the analysis of influences from the attribute sets, in above tables showing performance measures for different fractional factorial designs the conjoint model with main effects and without interaction terms showed overall best predictive performance over measures and measurement conditions. Therefore, value structure for all three comparisons concerning reliability over stimulus set are presented for the main effects conjoint model here, too.

Table XXXIII on page 174 shows scaled part-worths, importances, and importance ranks for the between-subjects comparison of groups G3 consisting of fractional factorial FF1, and G4 consisting of fractional factorial FF2. Once again, in agreement with consumer research literature, the three (3) most important attributes comprise over 50% of importance weights for all nine (9) attributes (53.3% for G3 and 54.1% for G4), i.e. the first three most important attributes explain over 50% of deviations in the response. Examining absolute importances, Base-Price, again, is the most important attribute for both fractional factorial conditions with 22.9% for G3 with fractional factorial FF1, and 26.6% for G4 with fractional factorial FF2. Second in attribute importance is Features with 18.1% for G3 and 16.3% for G4, and third in importance is Type-of-Display with 12.3% for G3, and Battery-Life with 11.2% for G4. Considering differences in importances, the most marked effect is the nonsignificant deviation in importance of price, the most important attribute, for the two conditions (difference of about 3.7%). Considering the other eight attributes, the difference between attribute importances is between one (1) and less than four (4) percentage points. This shows very high reliability over stimulus set for all attributes. However, some counter-intuitive deviation from expected level-utility functions occurs again for Screen-Size for FF1, but now also for FF2: For Screen-Size with

TABLE XXXIII

PART-WORTHS AND ATTRIBUTE IMPORTANCES G3 / G4 (TC MAIN EFFECTS)

Attribute Levels: Coded (Actual)	Scaled Part-Worths		Importance		Importance Rank
	G3	G4	G3	G4	G3 / G4
Weight-1 (9 pounds [a])	-0.8377	-0.8149			
Weight0 (7 pounds)	0.1852	0.3424			
Weight1 (5 pounds [a])	0.6525	0.4726	9.81%	8.21%	4 / 5
ScrSiz-1 (8.4 inch diagonal [a])	-0.1163	-0.1635			
ScrSiz0 (9.4 inch diagonal)	-0.2169	0.1843			
ScrSiz1 (10.4 inch diagonal[a])	0.3332	-0.0208	7.01%	6.91%	8 / 7
Display-1 (Monochrome)	-1.2959	-1.1564			
Display1 (Color)	1.2959	1.1564	12.31%	10.34%	3 / 4
B_Price-1 ($ 3500 [a])	-2.6418	-3.4388			
B_Price0 ($ 2500)	-0.0612	0.5952			
B_Price1 ($ 1500 [a])	2.7030	2.8436	22.90%	26.58%	1 / 1
Keyb_Siz-1 (Smaller than regular size)	-0.3570	-0.4735			
Keyb_Siz1 (Regular size)	0.3570	0.4735	5.61%	6.79%	9 / 8
BattLife-1 (3 hours[a])	-0.4536	-1.2538			
BattLife0 (5 hours)	-0.0034	0.3268			
BattLife1 (7 hours[a])	0.4571	0.9269	7.94%	11.24%	6 / 3
Speed-1 (Comfortable for word-processing)	-0.2665	-0.4159			
Speed1 (Fast for big spreadsheet and imaging)	0.2665	0.4159	7.23%	5.42%	7 / 9
Features-1 (No additional features[a])	-1.9367	-1.8277			
Features0 (Expansion slots for keyboard, monitor, others)	0.2758	0.0660			
Features1 (Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others[a])	1.6609	1.7617	18.09%	16.32%	2 / 2
PointDev-1 (Mouse [a])	0.1855	0.0066			
PointDev0 (Trackball)	-0.1454	0.2839			
PointDev1 (Trackpad or other[a])	-0.0401	-0.2905	9.08%	8.20%	5 / 6

[a] Levels used for the 2-level extreme design of the holdout product profiles.

FF1, the medium value of 9.4 inches is less preferred than the low value of 8.4 inches. In contrast to FF1, this counter-intuitive attribute preference does not occur for the according levels in G4, but for different levels in the attribute with FF2: There, level-utility for the medium value of 9.4 inches in Screen-Size is more preferred than the high value of 10.4 inches. This may suggest a spurious effect of fractional factorials when preferences are not very pronounced, i.e. when the attribute is relatively unimportant, or its importance is below chance levels (in this study at about 11%).

In order to gauge if value structure for between-subjects effects of fractional factorials FF1 and FF2 are caused by subject variation, within-subject comparisons of value structure are compiled in Tables XXXIV on page 176 and XXXV on page 177 for paired first and second measurements of groups G3 and G4. One remarkable observation is complete preservation of importance ranks for paired comparison of group G4, though the difference in importance of Price between two measurements is 6.2%. Absence of conspicuous shifts in importances suggests reliability over stimulus set for value structure, too.

TABLE XXXIV

PART-WORTHS AND ATTRIBUTE IMPORTANCES PART G1 / G3 (TC MAIN EFFECTS)

Attribute Levels: Coded (Actual)	Scaled Part-Worths		Importance		Importance Rank
	Part G1	G3	Part G1	G3	Part G1/G3
Weight-1 (9 pounds [a])	-0.9320	-0.8377			
Weight0 (7 pounds)	0.4333	0.1852			
Weight1 (5 pounds [a])	0.4986	0.6525	11.36%	9.81%	3 / 4
ScrSiz-1 (8.4 inch diagonal [a])	0.0056	-0.1163			
ScrSiz0 (9.4 inch diagonal)	-0.1446	-0.2169			
ScrSiz1 (10.4 inch diagonal[a])	0.1390	0.3332	7.06%	7.01%	8 / 8
Display-1 (Monochrome)	-0.9015	-1.2959			
Display1 (Color)	0.9015	1.2959	10.07%	12.31%	6 / 3
B_Price-1 ($ 3500 [a])	-2.1374	-2.6418			
B_Price0 ($ 2500)	0.0928	-0.0612			
B_Price1 ($ 1500 [a])	2.0446	2.7030	18.95%	22.90%	1 / 1
Keyb_Siz-1 (Smaller than regular size)	-0.0334	-0.3570			
Keyb_Siz1 (Regular size)	0.0334	0.3570	5.52%	5.61%	9 / 9
BattLife-1 (3 hours[a])	-0.7199	-0.4536			
BattLife0 (5 hours)	0.0045	-0.0034			
BattLife1 (7 hours[a])	0.7154	0.4571	10.34%	7.94%	5 / 6
Speed-1 (Comfortable for word-processing)	-0.4040	-0.2665			
Speed1 (Fast for big spreadsheet and imaging)	0.4040	0.2665	7.40%	7.23%	7 / 7
Features-1 (No additional features[a])	-2.1139	-1.9367			
Features0 (Expansion slots for keyboard, monitor, others)	0.5351	0.2758			
Features1 (Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others[a])	1.5788	1.6609	18.14%	18.09%	2 / 2
PointDev-1 (Mouse [a])	0.3647	0.1855			
PointDev0 (Trackball)	-0.1779	-0.1454			
PointDev1 (Trackpad or other[a])	-0.1868	-0.0401	11.14%	9.08%	4 / 5

[a] Levels used for the 2-level extreme design of the holdout product profiles.

TABLE XXXV

PART-WORTHS AND ATTRIBUTE IMPORTANCES PART G1 / G4 (TC MAIN EFFECTS)

Attribute Levels: Coded (Actual)	Scaled Part-Worths		Importance		Importance Rank
	Part G1	G4	Part G1	G4	Part G1/G4
Weight-1 (9 pounds [a])	-0.7481	-0.8149			
Weight0 (7 pounds)	0.2803	0.3424			
Weight1 (5 pounds [a])	0.4678	0.4726	9.85%	8.21%	5 / 5
ScrSiz-1 (8.4 inch diagonal [a])	-0.0292	-0.1635			
ScrSiz0 (9.4 inch diagonal)	-0.2486	0.1843			
ScrSiz1 (10.4 inch diagonal[a])	0.2778	-0.0208	7.15%	6.91%	7 / 7
Display-1 (Monochrome)	-1.3086	-1.1564			
Display1 (Color)	1.3086	1.1564	12.02%	10.34%	4 / 4
B_Price-1 ($ 3500 [a])	-2.3020	-3.4388			
B_Price0 ($ 2500)	0.2879	0.5952			
B_Price1 ($ 1500 [a])	2.0141	2.8436	20.34%	26.58%	1 / 1
Keyb_Siz-1 (Smaller than regular size)	-0.1378	-0.4735			
Keyb_Siz1 (Regular size)	0.1378	0.4735	4.87%	6.79%	8 / 8
BattLife-1 (3 hours[a])	-1.4176	-1.2538			
BattLife0 (5 hours)	0.0943	0.3268			
BattLife1 (7 hours[a])	1.3233	0.9269	12.93%	11.24%	3 / 3
Speed-1 (Comfortable for word-processing)	0.1736	-0.4159			
Speed1 (Fast for big spreadsheet and imaging)	-0.1736	0.4159	4.49%	5.42%	9 / 9
Features-1 (No additional features[a])	-2.2772	-1.8277			
Features0 (Expansion slots for keyboard, monitor, others)	0.2763	0.0660			
Features1 (Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others[a])	2.0009	1.7617	19.20%	16.32%	2 / 2
PointDev-1 (Mouse [a])	0.6737	0.0066			
PointDev0 (Trackball)	-0.1944	0.2839			
PointDev1 (Trackpad or other[a])	-0.4794	-0.2905	9.14%	8.20%	6 / 6

[a] Levels used for the 2-level extreme design of the holdout product profiles.

4.2.3 Reliability and Interaction of Conjoint Methodological Factors

Research Question # 3 .

How do type of attribute in the product profile and factorial design interact in their influence on customer value structure for different models ?

From the literature review, no indication about the direction of this interaction for predictive accuracy is obtained. One general suggestion is that differences due to several methodological variations should cancel out. This leads to the following hypothesis:

H_0: The interaction of differences in attribute set and specific fractional factorial design does not influence predictive performance.

H_A: The interaction of differences in attribute set and specific fractional factorial design does influence predictive performance.

In order to test this hypothesis, groups G4 and G5 are compared and tested on their performance measures as a between-subjects comparison. Comparisons of value structure are presented after the F-tests for performance measures.

Predictive Performance

Table XXXVI on page 180 gives an overview over different performance measures for between-subjects comparison of groups G4 and G5. This comparison gauges differences due to the confounded effects of different fractional factorial designs (FF2 and FF1, respectively), and different attribute sets A1 and A2. For the majority of performance measures (39) and model forms, prediction in group G5 is better than in group G4. Only for two correlations r_{xy} (TC main effects and TC iBxD) and the best

First-Hit measure (TC main effects) G4 shows better prediction. From the comparisons and tests of the separate effects of attribute set and factorial design it may be concluded that effects from the type of attribute set, specifically from including the user-referent attribute Firm-Reputation into attribute set A2, is more determinant for an improvement in prediction than the possible deterioration from inclusion of factorial design FF2. This is exactly what can be observed in the overview of measures. Though the improvement in prediction with combination A2FF1 in G5 vs. combination A1FF2 in G4 is rather consistent, associated F-tests cannot establish significance of differences: All differences are clearly insignificant (cp. Table XXXVII on page 181). This may confirm the belief that interaction of influences from variation in attribute sets and fractional factorial design, and possibly other methodological variations, cancel out in their effect. A noteworthy observation is that higher R^2 of the holdout set of profiles, i.e. more variance explained for the validation set, is associated with lower (i.e. better) RMSE and higher Fisher's z, but also with lower (i.e. worse) correlation coefficient and lower First-Hit for the best conjoint model (TC main effects). This may show some different capability of measures to reflect the level of accuracy in prediction reached with the model, specifically when this level is high. It may be speculated that a high level of predictive accuracy may more easily show deviation from an interval scale for correlation r_{xy} , and distortions from some ill understood properties of First-Hit.

Nevertheless, from observation of results obtained for comparison of groups G4 and G5 one cannot reject the null hypothesis H_0 that the interaction of differences in attribute set and specific fractional factorial design does not influence predictive performance. In conclusion, conjoint analysis may be viewed as reliable over the conjoint effects of different attribute and stimulus sets.

TABLE XXXVI

PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS; G4 / G5

	Type of Model	Performance Measures (Averages Over Groups)								
		R^2 (calib.)	Adj R^2 (calib.)	R^2 (hold.) [b]	r_{xy} (hold.) [a]	Fisher's $z(r_{xy})$ (hold.) [a]	RMSE (hold.)	RMSE (calib.)	First-Hit (hold.)	First-Hit (mean counts)
1.	TC main effects	0.8548 / 0.8693	0.6569 / 0.6910	0.5963 / 0.6003	0.7615 / 0.7468	1.0672 / 1.0998	20.11 / 19.35	8.54 / 7.24	74.17% / 70.83%	2.97 / 2.83
2.	TC iAxD	0.8778 / 0.8871	0.6470 / 0.6737	0.5650 / 0.5910	0.7379 / 0.7448	1.0179 / 1.0809	21.00 / 19.69	7.80 / 6.75	71.67% / 73.33%	2.87 / 2.93
3.	TC iBxD	0.8767 / 0.8918	0.6438 / 0.6875	0.5013 / 0.5211	0.6896 / 0.6881	0.9182 / 0.9575	23.40 / 22.21	7.83 / 6.37	61.67% / 69.17%	2.47 / 2.77
4.	TC iCxD	0.8821 / 0.8909	0.6593 / 0.6849	0.5478 / 0.5900	0.7236 / 0.7391	0.9960 / 1.0792	21.59 / 19.80	7.56 / 6.62	70.00% / 70.83%	2.80 / 2.83
5.	SE	n/a	n/a	0.4825 / 0.5307	0.6565 / 0.7104	0.8794 / 0.9573	29.00 / 26.14	n/a	67.50% / 70.83%	2.70 / 2.83

(calib.) = Calibration set
(hold.) = Holdout set

[a] = Seemingly non-monotone transformations between r_{xy} and $z(r_{xy})$ when comparing different cells in the table result from averaging individual results which is appropriate for Fisher's z, but not for r_{xy}.

[b] = Averaged from the individual R^2s (calculations from group r_{xy} show too low coefficients).

TABLE XXXVII

F-TESTS OF PREDICTIVE PERFORMANCES OF INDIVIDUAL-LEVEL MODELS (G4 / G5 ; $F_{1,58}$ DEGREES OF FREEDOM)

Type of Model	F-Tests of Performance Measures (Averages Over Groups)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G4 / G5	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G4 / G5	$F_{1,58}$ [a] (p-Value) R^2	Conf. Int. (90%) [b] G4 / G5
1. TC main effects	0.1076 (.7441) 0.19%	[0.9731;1.1613] / [0.9597;1.2399]	0.1952 (.6603) 0.34%	[18.10;22.13] / [17.21;21.49]	0.3760 (.5421) 0.64%	[2.72;3.22] / [2.56;3.10]
2. TC iAxD	0.4226 (.5182) 0.72%	[0.9184;1.1175] / [0.9498;1.2120]	0.6093 (.4382) 1.04%	[19.10;22.91] / [17.58;21.81]	0.0817 (.7760) 0.14%	[2.56;3.17] / [2.68;3.19]
3. TC iBxD	0.1494 (.7006) 0.26%	[0.8142;1.0222] / [0.8195;1.0955]	0.3258 (.5703) 0.56%	[21.13;25.68] / [19.49;24.93]	1.6034 (.2105) 2.69%	[2.21;2.72] / [2.45;3.08]
4. TC iCxD	0.6601 (.4199) 1.13%	[0.8891;1.1028] / [0.9417;1.2168]	0.9577 (.3318) 1.62%	[19.37;23.81] / [17.63;21.98]	0.0215 (.8839) 0.04%	[2.54;3.06] / [2.55;3.12]
5. SE	0.6988 (.4066) 1.19%	[0.7572;1.0017] / [0.8568;1.0579]	1.0574 (.3081) 1.79%	[26.25;31.75] / [22.30;29.98]	0.2195 (.6412) 0.38%	[2.31;3.09] / [2.55;3.12]

(hold.) = Holdout set; [a] = Set of both groups; [b] = Group forms its own set (two-tailed, $\alpha = 0.1$; DFs 29/29)

Value Structure

Comparing value structure for different methodological variations in terms of attribute sets and fractional factorial designs included, the conjoint model with main effects and without interaction terms showed overall best predictive performance over measures and measurement conditions. This model´s value structure is presented below in Table XXXVIII on page 183.

The interaction of attribute set and fractional factorial included in the study design shows deviations in importances between one (1) and less than three (3) percentage points. This is even less than observed with different fractional factorial designs. Still, Base-Price and Features are the two most important attributes in both groups (26.6% and 16.3% in group G4, and 28.8% and 15.4% in group G5). The third most important attribute is Battery-Life for G4 with 11.2% in importance, and Firm-Reputation for G5 with 10.5% in importance. In both cases, these three attributes comprise over 50% of the importances, i.e. explain over 50% of deviations (54.1% for G4 and 54.7% for G5). Some counter-intuitive level-utility functions occur as already discussed in prior sections. In sum, however, value structure may safely be regarded as reliable over the interaction of variations in attribute set and fractional factorial design.

TABLE XXXVIII

PART-WORTHS AND ATTRIBUTE IMPORTANCES G4 / G5 (TC MAIN EFFECTS)

Attribute Levels: Coded (Actual)	Scaled Part-Worths		Importance		Importance Rank
	G4	G5	G4	G5	G4 / G5
Weight-1 (9 pounds [a])	-0.8149	-0.3505			
Weight0 (7 pounds)	0.3424	0.2902			
Weight1 (5 pounds [a])	0.4726	0.0604	8.21%	7.34%	5 / 7
ScrSiz-1 (8.4 inch diagonal [a])	-0.1635	-0.0154			
ScrSiz0 (9.4 inch diagonal)	0.1843	-0.2036			
ScrSiz1 (10.4 inch diagonal[a])	-0.0208	0.2190	6.91%	7.47%	7 / 6
Display-1 (Monochrome)	-1.1564	-1.1342			
Display1 (Color)	1.1564	1.1342	10.34%	10.22%	4 / 4
B_Price-1 ($ 3500 [a])	-3.4388	-3.4202			
B_Price0 ($ 2500)	0.5952	0.2849			
B_Price1 ($ 1500 [a])	2.8436	3.1353	26.58%	28.84%	1 / 1
Keyb_Siz-1 (Smaller than regular size)	-0.4735	-0.0211			
Keyb_Siz1 (Regular size)	0.4735	0.0211	6.79%	4.37%	8 / 9
BattLife-1 (3 hours[a])	-1.2538	-0.0962			
BattLife0 (5 hours)	0.3268	-0.2178			
BattLife1 (7 hours[a])	0.9269	0.3140	11.24%	9.17%	3 / 5
Speed-1 (Comfortable for word-processing)	-0.4159	-0.1576			
Speed1 (Fast for big spreadsheet and imaging)	0.4159	0.1576	5.42%	6.71%	9 / 8
Features-1 (No additional features[a])	-1.8277	-1.6968			
Features0 (Expansion slots for keyboard, monitor, others)	0.0660	0.5716			
Features1 (Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others[a])	1.7617	1.1252	16.32%	15.37%	2 / 2
PointDev-1 / Firm_Rep-1	0.0066	-0.7544			
PointDev0 / Firm_Rep0	0.2839	-0.0278			
PointDev1 / Firm_Rep1	-0.2905	0.7822	8.20%	10.52%	6 / 3

[a] Levels used for the 2-level extreme design of the holdout product profiles.

4.2.4 Relative Performance of Individual-Level Models

Research Question # 4.

Which individual-level model for customer value structure performs best with respect to prediction ?

As has already been stated in section 3.3.6 on pp. 125, from the literature review, the only indication about the direction of relative performance of individual-level models is suggested superiority of (traditional; TC) conjoint models over self-explicated (SE) models. However, for methodological variations and a variety of situations no general statements about predictive accuracy of models with interactions and without them were obtained. This leads to the following hypothesis:

H_0: Individual-level models for customer value structure do not distinguish themselves in terms of predictive performance.

H_A: Individual-level models for customer value structure do distinguish themselves in terms of predictive performance.

This hypothesis is tested using multi-way ANOVAs for performance measures and the five types of models. The tests are performed with all 2nd group estimates and selected performance measures (Fisher´s z-transformed correlation coefficients, RMSE, and First-Hit), as a Student´s t-test for each pair of model forms, and individual comparisons only (Table XXXIX on page 187). Where higher than the ANOVAs´ accuracy is needed to determine significance of differences between individual-level models, paired t-tests are performed for Fisher´s z and RMSE measures, and for selected paired comparisons of model forms.

Judgment Criteria and Measures

Determination of "best" individual-level model takes into account the criteria

- performance in absolute terms with respect to objectives, i.e.
    - accuracy in prediction,
    - substantiality of value structure for segmentation, and
- relative performance with respect to parsimony of models.

Measures to judge relative performance of models are not easy to determine, as this decision depends on the objective pursued. If the objective is highest predictive accuracy, then an absolute or incremental performance measure is appropriate, as for instance First-Hit and R^2. However, when parsimony of the model shall be taken into account, a parsimonious performance measure, i.e. one that takes the number of model parameters into account, is more appropriate.

Judgment of relative performance of conjoint models also concerns the issue of increased performance with more parameters, i.e. with interaction terms, vs. worse performance because of decreased degrees of freedom, increasing bias in estimation.

A practical consideration not to be neglected is the availability of tests when considering choice of appropriate measures for comparisons between models. In this study this problem arises as First-Hit is testable on the group level and over all respondents, but does not satisfy the assumption of normal distribution of responses. Also, RMSE shows a nonnormal (logistic) distribution over all forms of models and respondents´ measurements. Even Fisher´s z shows marginal deviation from normality[22], nevertheless all three measures were used to test significance of

[22] This is in contrast to the tests conducted for specific methodological groups where the assumption of normal distribution of responses could not be rejected, and it illustrates the fact that just by increasing the number of responses, significance is detected even for minor differences.

differences between means of model forms in order to gauge consequences of violations of test assumptions on the t-tests between pairs of models, especially in the case of First-Hit measure. This is relevant for tests of segment-level models, when the only measure to be tested is First-Hit. The ANOVAs with Fisher´s z and RMSE, respectively, provide the more reliable tests. For selected Fisher´s z and RMSE measures a paired t-test was performed.

Predictive Performance

Results of tests are presented in Tables XXXIX and XL on pages 187 and 188. Table XXXIX consistently shows TC main effects model as the one with the best mean performance over all respondents and methodological variations, i.e. the highest (Fisher´s z-transformed) correlations between actual and predicted holdout evaluation, the highest First-Hit, and the lowest RMSE. In Table XL tests show consistent results, i.e. no significance of differences between models for the three (3) best models (TC main effects, TC iAxD, TC iCxD), over all three measures, and for different model forms, but show test differences for the worst two models (TC iBxD, SE). However, tests cannot confirm significance of differences between TC main effects and the two second best models (TC iAxD, TC iCxD). Differences between TC iBxD and SE models are only significant with RMSE but not with Fisher´s z and First-Hit measures. Therefore, and in addition to above tests, three paired t-tests between TC main effects model and TC iAxD / TC iCxD, as well as between TC iBxD and SE models were conducted with Fisher´s z-transformed correlations and RMSE, as paired t-tests provide for stronger tests when the assumption of normal distribution is not violated. Paired t-tests between these selected modelscould determine significance of differences. Table XLI on page 189 provides for a summary of the results.

Independent from statistical significance, the researcher should determine substantial relevance of the magnitudes of differences.

TABLE XXXIX

RELATIVE PERFORMANCE OF INDIVIDUAL-LEVEL SELF-EXPLICATED AND CONJOINT MODEL FORMS

Levels	F-Tests of Performance Measures (Averages Over All 117 Respondents)					
	Fisher's z (r_{xy}) (hold.)		RMSE (hold.)		First-Hit (mean counts)	
(Type of Model)	$F_{4.580}$ [a] (p-Value)	(Power α = 0.05) R^2 (Power α = 0.1)	$F_{4.580}$ [a] (p-Value)	(Power α = 0.05) R^2 (Power α = 0.1)	$F_{4.580}$ [a] (p-Value)	(Power α = 0.05) R^2 (Power α = 0.1)
	5.9600 (0.0001)	(0.9851) 0.0395 (0.9935)	16.5798 (0.0000)	(1.0000) 0.1026 (1.0000)	2.7437 (0.0278)	(1.0000) 0.0186 (1.0000)
	Mean [b]	Std. Deviation [b]	Mean [b]	Std. Deviation [b]	Mean [b]	Std. Deviation [b]
1. TC main effects	1.0410	0.0354	19.6885	0.6815	2.8205	0.8671
2. TC iAxD	0.9965	0.0352	20.3820	0.6879	2.8205	0.8965
3. TC iBxD	0.8717	0.0355	23.9500	0.9000	2.4957	0.9615
4. TC iCxD	1.0015	0.0357	20.3408	0.6971	2.7692	0.8846
5. SE	0.8533	0.0308	27.5312	1.0551	2.6239	1.0316

(hold.) = Holdout set; [a] = Set of all groups (levels); [b] = Group forms its own set

TABLE XL

PAIRED (STUDENT'S) T-TESTS OF INDIVIDUAL-LEVEL MODEL FORMS

Fisher's z(r_{xy})

t = 1.9641
Alpha = 0.05

Abs(Dif)-LSD	TC main	TC iCxD	TC iAxD	TC iBxD	SE
TC main	-0.0961	-0.0566	-0.0516	0.0731	0.0916
TC iCxD	-0.0566	-0.0961	-0.0911	0.0337	0.0521
TC iAxD	-0.0516	-0.0911	-0.0961	0.0287	0.0471
TC iBxD	0.0731	0.0337	0.0287	-0.0961	-0.0777
SE	0.0916	0.0521	0.0471	-0.0777	-0.0961

RMSE

t = 1.9641
Alpha = 0.05

Abs(Dif)-LSD	SE	TC iBxD	TC iAxD	TC iCxD	TC main
SE	-2.2726	1.3086	4.8766	4.9178	5.5701
TC iBxD	1.3086	-2.2726	1.2955	1.3366	1.9889
TC iAxD	4.8766	1.2955	-2.2726	-2.2314	-1.5792
TC iCxD	4.9178	1.3366	-2.2314	-2.2726	-1.6204
TC main	5.5701	1.9889	-1.5792	-1.6204	-2.2726

First-Hit

t = 1.6475
Alpha = 0.1

Abs(Dif)-LSD	TC main	TC iAxD	TC iCxD	SE	TC iBxD
TC main	-0.2004	-0.2004	-0.1491	-0.0038	0.1244
TC iAxD	-0.2004	-0.2004	-0.1491	-0.0038	0.1244
TC iCxD	-0.1491	-0.1491	-0.2004	-0.0551	0.0731
SE	-0.0038	-0.0038	-0.0551	-0.2004	-0.0722
TC iBxD	0.1244	0.1244	0.0731	-0.0722	-0.2004

Abs(Dif)-LSD = Absolute difference to the overall mean minus the least significant difference. Thus, positive values show pairs of means that are significantly different.
Comparisons are for each pair using Student's t.
Rows are ordered according to increasing magnitude of differences between models with the first column.

TABLE XLI

SELECTED PAIRED T-TESTS OF THE THREE(3) BEST-PERFORMING, AND THE TWO (2) WORST-PERFORMING MODEL FORMS

Paired t-tests (one-tailed) Model 1 vs. Model 2	Performance Measure	Assumption of normal distribution valid ?	t-value	p-value
TC main effects vs.	Fisher's z	Yes	3.9746	0.0001
TC iAxD	RMSE	Yes	4.3270	0.0000
TC main effects vs.	Fisher's z	Yes	4.0339	0.0000
TC ;CxD	RMSE	Yes	4.6825	0.0000
TC iBxD vs.	Fisher's z	Yes	0.4923	0.3117
SE	RMSE	No (marginally)	3.3985	0.0005

With performances of different individual-level model forms in Table XXXIX on page 187, Student's t-tests in Table XL on page 188, and paired t-tests in Table XLI on page 189 the comparison of models yields clear results: Best model in terms of performance measures for accuracy in prediction, and confirmed with multi-way ANOVAs and paired t-tests, is the traditional conjoint model with main effects only (TC main effects). Worst model over all responses for Fisher's z and RMSE, and second worst for First-Hit is the self-explicated model (SE). Tests of significance of this difference, however, are inconsistent: While RMSE detects a significant difference between mean performances of these two models, Fisher's z does not. Those findings confirm assumed superiority of conjoint models over self-explicated models. As for conjoint models with interaction terms, performance is dependent on the interaction modeled. These models can be among the best and among the worst.

In consequence, and summarizing these findings, one can reject H_0 that individual-level models for customer value structure do not distinguish themselves in terms of predictive performance. Therefore, H_A must be believed, i.e. that individual-level models for customer value structure do distinguish themselves in accuracy of prediction. The traditional conjoint model with main effects only (TC main effects) is the best overall model. The self-explicated model (SE) is (among) the worst one(s).

4.2.5 Summary of Results in Phase I

Phase I of this research study revealed four (4) major findings. First, conjoint analysis is reliable over the attribute set chosen. However, there is a tendency of user-referent attributes to increase predictive accuracy, though this finding could not be confirmed unambiguously with appropriate tests. Second, conjoint analysis may safely be regarded as reliable over the stimulus set: No effect, whatsoever, could be detected. Third, the interaction of changes in the attribute set and stimulus set does not influence external reliability of conjoint models. The conjecture that effects of methodological variations do cancel out and do not add up seems to hold. Fouth, the best model in terms of accuracy in prediction of preferences and choice behavior is the traditional conjoint model with main effects only. Another main finding is reliability of conjoint models over time. However, it seems that accuracy in measurement may be increased by simply measuring twice.

As the traditional conjoint model with main effects only was found best predictive model, it is used in Phase II of this study to explore possible improvements in predictive accuracy with segment-level conjoint models. Phase II aims at an empirical validation of Hagerty´s claim (1986, p. 301 and p. 309) that a reversal of the best conjoint model is probable with a change from individual to market conjoint models.

4.3 Phase II

Rationales for respondent grouping with three methods (HIC, NHC, FUC) and clustering in connection with conjoint analysis have been provided in section 2.5 of this study. Here, three (3) different clustering procedures (HIC, NHC, FUC) are applied to subjects exhibiting different benefit attributions to product profiles, i.e. part-worth utilities. Some general remarks about pattern recognition with clustering and provisions for comparability precede rationales for choices of clustering parameters. Results of clustering procedures followed by conjoint results for the three grouping methods are presented, next. Finally, segment-level results are compared to the individual-level results in terms of prediction and value structure.

4.3.1 Supra-Level Perspective of Segmentation With Clustering Procedures

Segmentation with clustering procedures may be viewed as part of the general problem of pattern recognition as a "search for structure in data" (Bezdek 1981, p. 1). A prerequisite and presumption for the search to be successful is that data carry information about the process generating it. This is an issue of variable and feature selection for the search procedures. The type of search performed depends not only on the data and our models, but upon the structure we expect to find. Structure, here, means there is a way to organize information from the data in a manner that exposes relationships between variables in the process, i.e. product attributes and preferences or choice behavior. As a representation of structure conveys specific types and amounts of information, one may express the elements of pattern recognition in terms of information as, "the data contain it, the search recognizes it, and the structure represents it" (Bezdek 1981, p. 2).

Notions of information, precision, and usefulness are central for understanding of segmentation, or pattern recognition in general[23]. These three notions are closely related but also exhibit a certain tension when pursued, ensemble. The motivation behind segmentation in marketing, and specifically in conjoint analysis, and the criticality of segmentation base and segmentation method for identification of segments has been exposed in section 2.5 of this study. Segmentation is useful when parts of the market show commonality in their preferences or market behavior that is distinct from other parts of the market (substantiality[24]), and this commonality may be linked to, or may be influenced with managerial actions which emanate from business objectives (accessibility and actionability). Different objectives, however, usually necessitate different types of information, yielding different levels of precision when information is measured with one specific method. The increase in precision of information for one type of information, satisfying one objective, may decrease precision of information for other types of information, satisfying other objectives.

This is the type of problem encountered in conjoint analysis when pursuing objectives of prediction and segmentation. Clustering procedures have been advanced to supposedly improve on both objectives: Increasing reliability of parameter estimates, i.e. of value structure, by trading high variance in respondents´ part-worth estimates for increased parameter stability (i.e. less bias in part-worth estimation). This may be useful for prediction. However, segmentation and structural identification of markets may be better served with less variance in the respondents which comes with increased bias in part-worth estimates, as parameters are derived individually for each

[23] The ideas presented are heavily influenced by the teachings of George Klir, and his readings as well as those of Lotfi Zadeh, James Bezdek, and Bart Kosko. The latter expressed the conflict between more information and precision as "information up, fuzz up" (Presentation in Portland, OR, February 17, 1995).

[24] Substantiality comprises not only discriminatory behaviors but also their stability over time.

respondent. Thus, model form and other methodological choices (not only in conjoint analysis) have a direct bearing on reaching respective objectives, as well as on the precision with which they may be reached.

Traditionally, segmentation bases have been chosen with managerial judgment from experience, field studies, and other sources of information about the distinct features of market participants, linking structural information about customers, e.g. their demographics, to their behavior in the marketplace. This approach is very imprecise and highly subjective. Automation of the "search" promises to find opaque, or non-intuitive patterns, as well as being more objective about the potential of features to covary with preferences and market behavior. The search need not be performed with cluster algorithms, but may be performed with other methods as well, as for instance with information theory (Hosseini 1987). The general purpose of using cluster analysis is to distill, i.e. identify, "natural" groupings of data through an automated, objective mechanism, i.e. search procedure. It is guaranteed, at least for the algorithmic procedures used in this study (HIC, NHC, and FUC), that the members of each cluster found with some well-defined operation are more similar to one another than to members of other clusters. At least, this is true in some mathematical sense, but one hopes that the same substructure exists in the data-generating process itself, being able to interpret cluster solutions in a useful manner. Therefore, a note of caution about the potential to "automatically" cluster data may be replicated, here:

"In view of the ... above, it is clear that *successful* cluster analysis ultimately rests not with the computer, but with the investigator, who is well advised to use some empirical hindsight concerning the physical process generating X [the matrix of observations; explanation added by author] to temper algorithmically suggested solutions. Specification of a similarity measure and/or clustering criterion is not enough. The method used must be matched to the data ... We reiterate that different similarity measures, clustering criteria, and axiomatic structures lead to astonishingly disparate structural interpretations of the same data set." (Bezdek 1981, p. 45 and p. 47)

These cautionary notes may be irrelevant when the sole purpose of clustering is reduction of dimensionality of feature space. This, however, leaves the question of what resulting objects the analyst is operating on, i.e. if those clusters do have any interpretation in reality, rather than as abstract objects of mathematics.

Partitioning the data, i.e. the process of determining whether and how clusters may be formed, involves four (4) major questions:

1. What variables or features should be used in computing similarity among objects ?
2. How should similarity be measured ?
3. What procedure or algorithm should be used for grouping, i.e. clustering ?
4. How many clusters should be formed (cluster validity) ?

There is no definitive answer to these four (4) questions, and no "right" approach, no single answer (Hair, Anderson, Tatham, and Black 1992, p. 270). One may view cluster analysis, just as factor analysis, more as an art than a science. The essential criterion for partitioning the data is to maximize differences among clusters relative to the variation within clusters. The choices made in this study are detailed in the following section.

4.3.2 Choice of Clustering Parameters

All or selected features as cluster base ?

Basis for clustering in connection with conjoint analysis are benefit attributions of respondents to product profiles. As there is no explicit theory providing a rationale for variable selection or choice of the number of clusters, i.e. if all part-worth utilities should be used or only selected ones, and if few or many clusters should be allowed, supposition, past research, and practical considerations serve as guides in this process.

From experience and past research one may suggest different types of buyers for the study´s measurement object, a notebook or laptop computer:

- One type of potential buyers may be characterized by their desire to own the best and most recent product, putting high value on the latest features and technical possibilities. In terms of product adoption dynamics in new markets these may be termed the innovators or the early adopters. They are often knowledgeable about the technical possibilities and already familiar with the concepts applied in the product. Experience with the product class, or special needs for the product´s benefits may also increase benefit attributions to advanced features, i.e. feature-sensitivity. These buyers may also be termed the optimalists.
- A different type of potential buyers may primarily be characterized by their price-consciousness. They seek product benefits only after careful deliberation, and comparison with what they have to give up to obtain these benefits. New features have to work, and do not justify much more additional monetary sacrifice. They may be termed the minimalists.
- A third type of potential buyers in this study may be characterized as respondents having pronounced preference structure for specific features, whereas others may not have distinct preferences. These types of respondents may be termed categorists and averagers, respectively. Together with respective features, this may provide for yet another line of delineation of groups for marked differences in benefit attributions to product features.

Further profiling, having clustered the data, may provide similar or more appropriate lines of delineation and according labels for describing the characteristics of the clusters obtained. However, the main interest in this study is in the cluster algorithms´ potential to increase accuracy of measurement of benefit attributions to profile descriptions. Therefore, in order to allow any combination of benefit attribution to come to bear and not restrict algorithms to find only preconceived groups, all part-worths, i.e. attribute level utilities, except for the perturbed attributes, are used as input search space for the cluster algorithms. Deliberations about the appropriateness to pool respondents for clustering despite methodological variations are provided at the end of this section on pp. 201.

Number of clusters

Deliberations about the type of potential buyers are also relevant for determining the number of clusters, as clusters must be interpretable in terms of the research area if they are not to be treated as algorithmic artifacts without substantive meaning. Hierarchical clustering (HIC) and fuzzy clustering (FUC) provide some suggestions as to the appropriate number of clusters based on their mathematical properties. The former allows for a scree test using subsequent increase in distance measure as basis for deciding upon the number of clusters. The latter suggests use of allowed overlap among the ranges of the feature space, i.e. overlap of ranges of part-worth utilities, for potential cluster centers as basis for deciding upon the number of clusters. Details are provided in respective sections and the literature.

Another deliberation about the number of clusters concerns outliers and noisy data: Those respondents without distinct preferences are either randomly falling within or near a cluster, or they lie in a distance from any cluster. Usually, cluster algorithms do not allow the option to qualify the latter responses as "no cluster points". Therefore, with some cluster algorithms, one or a few unfortunately distributed outliers can severely distort results in terms of the number of clusters found and their respective centers. Possible precautions against outliers and influential observations are dependent on the algorithm applied. The fuzzy cluster algorithm applied here allows to specify acceptance of a data point for a cluster based on the point´s membership value. This is one way to specify a "range of influence" of a cluster center on its neighboring points, and thus exclude "outliers".

Another way to identify outliers is possible with agglomerative algorithms that are based on (hyperspherical) nearness between the actual points (not their respective cluster centers): Outliers and influential observations are added late in the process, i.e. they form their own clusters until late in the process while most other points have already grouped together. These former points´ distances to initial ("real") clusters is high, identifying them as outliers and influential observations. Unfortunately, the algorithms applied do not allow to specify a criterion to stop the search for structure when for instance 95% of the data has been clustered.

The previous two paragraphs illustrated decisions about the number of clusters based on the exclusion of noisy data. Additionally, while considerations about the types of potential buyers suggests at least three (3) clusters, the sample size of 117 respondents suggests a limit of about 5 to 6 meaningful clusters. Finally, the number of clusters should be about equal across the clustering procedures used in this study in order to provide for a "fair" comparison of the performance of cluster algorithms on their potential for increased accuracy in conjoint measurement. This requirement, however, conflicts with the desire to bring a specific method´s full potential to bear. After exploratory trials, and after examining a scree test, this question was resolved by generally considering three (3) and four (4) clusters for each cluster algorithm.

Scale of the cluster base feature vector

The conjoint literature that uses clustering is very divergent when deciding upon cluster bases and their scales: Regression coefficients (standardized or from their original scales), scaled part-worths, and importances were all used as cluster bases. Using the regression coefficients directly in either form, does not seem to be appropriate, first, because they do not represent all the part-worth utilities directly but as an, at most, m*(i-1) subspace (cp. equations E3.1.4 to E3.1.7 on pages 81 and 82), and second, because they are not comparable across respondents due to individual estimations and individual response patterns, i.e. unequal variances.

The most common recommendation to remedy comparability across respondents is to standardize or to normalize the input data. Though this remedies inter-subject comparability on single part-worths, it introduces new bias because the relations between different part-worth utilities within the individual are distorted. Conjoint part-worth utilities are not independent from each other. As data transformation introduces its own bias by putting constraints on the data (e.g. forcing values to map between 0 and 1), cluster base data, i.e. part-worth utilities, should remain as close as possible to the original data[25].

Furthermore, clustering must be based on the discriminatory features while allowing for fair inter-subject comparison[26]. Importances derived from the ranges of attribute utilities seem to have the desired properties for comparison: they are already normed, and they are directly interpretable as the discriminating elements. However, they do

[25] Overall (profile) utilities are closest to the original decision context, but they provide no discriminatory information with respect to attribute and level influences.

[26] The discriminating information is in the ranges (i.e. importances), not in the level utilities per se. One may also say it is in the level utilities *and* the intercept, i.e. part-worths may be interpreted as deviations from the mean response: the larger the range of deviation, the larger the relative influence of the attribute, the higher the attribute's importance.

not provide information about the utility or disutility of individual attribute levels, only about relative influences of all levels of an attribute on the total preference of a profile. Thus, from the above follows, this study uses all attribute level utilities of an individual, except for the perturbed ones, scaled with the sum of the ranges over all levels of all attributes according to the following formula[27]:

$$\text{Scaled } \alpha_{ij} = \frac{(L_j * m) * \alpha_{ij}}{\sum_{j=1}^{m} [\underset{i}{\mathrm{Max}}(\alpha_{ij}) - \underset{i}{\mathrm{Min}}(\alpha_{ij});\ \ ]_{i}} \quad \text{, for each ij} \qquad \text{(E5.1)}$$

where $(L_j * m)$ simply indicates the total number of levels in the model, and $[\underset{i}{\mathrm{Max}}(\alpha_{ij}) - \underset{i}{\mathrm{Min}}(\alpha_{ij})]$ denotes the range of part-worths over all levels i of attribute j.

This kind of scaling of the original part-worths makes them comparable between respondents but preserves relations among part-worths and their ranges (importances) within subjects.

Similarity measure

There are basically two types of distance or similarity measures used in cluster algorithms: First, measures based on Euclidean distance which uses squared differences, second, measures based on absolute or city-block distance which uses the sum of the absolute differences, both with respective adjustments for different kinds of situations. This study uses only algorithms with Euclidean distance measures.

[27] Multiplying this value with the number of levels in the model (L_j*m, here 24) allows also comparison over models with varying numbers of model parameters, e.g. with inclusion of interaction terms.

Admittedly, these measures are prone to find primarily hyperspherical cluster shapes which may not represent the data adequately.

Clustering algorithm

Due to the dimensionality of the feature space in this study (21),[28] and bearing in mind the prior paragraph, there is no way to determine if graph-based or objective-function-based algorithms are more appropriate for the data. Graph-based algorithms, like single-linkage, are more appropriate for data with "chains" or non-convex structures. Their disadvantage is often lack of generating a representative of each cluster (Bezdek 1981, p. 46). Objective-function-based algorithms are most appropriate for data which are basically hyperspherical and of roughly equal proportions (Bezdek 1981, p. 47). In exploratory analyses of two-dimensional slices through the data, no chain structures could be detected. Therefore, objective-function-based algorithms with Euclidean distance metrics are used in this study.

Which administration and which responses to cluster ?

As is revealed with Phase I of the study, though often not significant, the second measurement seems to be more accurate in prediction over most methodological variations, specifically it seems to explain more variance with the holdout set of data than the first measurement. Task familiarity obviously serves to reduce error which is administration-based. The second measurement seems to lead to more stable preferences without leading to a learning effect with respect to the responses for profiles. Therefore, the second measurement is used for clustering.

[28] The number twenty-one (21) is the number of attribute levels for the main-effects conjoint model (24) minus the number of levels of the perturbed attributes which are no bases for clustering.

Pooling of data

As for the methodological variations applied to this study's conjoint measurement, there may be a concern if it is appropriate to cluster data which has been gathered under different methodological variations. However, Phase I of this study could not determine significant effects of methodological variations (i.e. different types of attribute sets and different fractional factorial designs) on prediction and value structure, except for one increase in the importance of price when a user-referent attribute was introduced into the set of profiles. But even this one instance seems to be a spurious effect rather than a systematic one.

Specifically, pooling respondents that were administered different fractional factorials FF1 and FF2 is justified, as there is no change in substance concerning the attributes. Neither did results in Phase I for prediction detect group differences based on the fractional factorials (and thus also none of significance), nor did results for value structure, i.e. part-worths and importances. Moreover, pooling of respondents who received different attribute sets A1 and A2 is appropriate in this study, though in general it is not. Results of Phase I indicated no substantial shifts in accuracy of prediction, and for the most part none of significance. Though it may not necessarily be assumed that respondents with attribute set A2 are falling randomly within different clusters, there is also no indication of the opposite. Finally, it does not seem adequate to first cluster groups delineated by their methodological variations and then merge respective groups, as the character of clusters found usually is not preserved when the same procedures are applied to different groups. In conclusion, it is appropriate to pool all 117 responses and cluster them according to their value attributions to product features.

Cluster validity

Some additional deliberations about cluster validity apart from aforementioned considerations about classes of buyers and the number of clusters shall be made at this point. There are basically three (3) criteria to judge cluster validity:

1. Are the partitions obtained substantively interpretable ?
2. Do conjectured clusters exhibit distinct differences ?
3. Do different cluster methods arrive at similar partitions ?

Cf. 1:This question may be answered, mainly, with the discussion about classes of buyers and the resulting number of clusters on page 195. It is the most important question for judgments about cluster validity.

Cf. 2:This question may be answered by plotting scaled part-worth utilities of cluster centers for partitions obtained with different algorithms on top of each other. This contrasts those differences, and allows for substantive interpretation. Figures 10 to 23 on pages 203 to 209 exhibit cluster profiles for all fourteen (14) cluster segmentation procedures applied in this study.

Cf. 3:If this question can be answered positively, it serves as an additional criterion to increase belief in algorithmic cluster solutions. Tables XLII on page 210 and XLIII on page 211 allow for checks of relative compatibility among clusters obtained with different algorithms, on the basis of percentages, and on the basis of actual numbers of overlap.

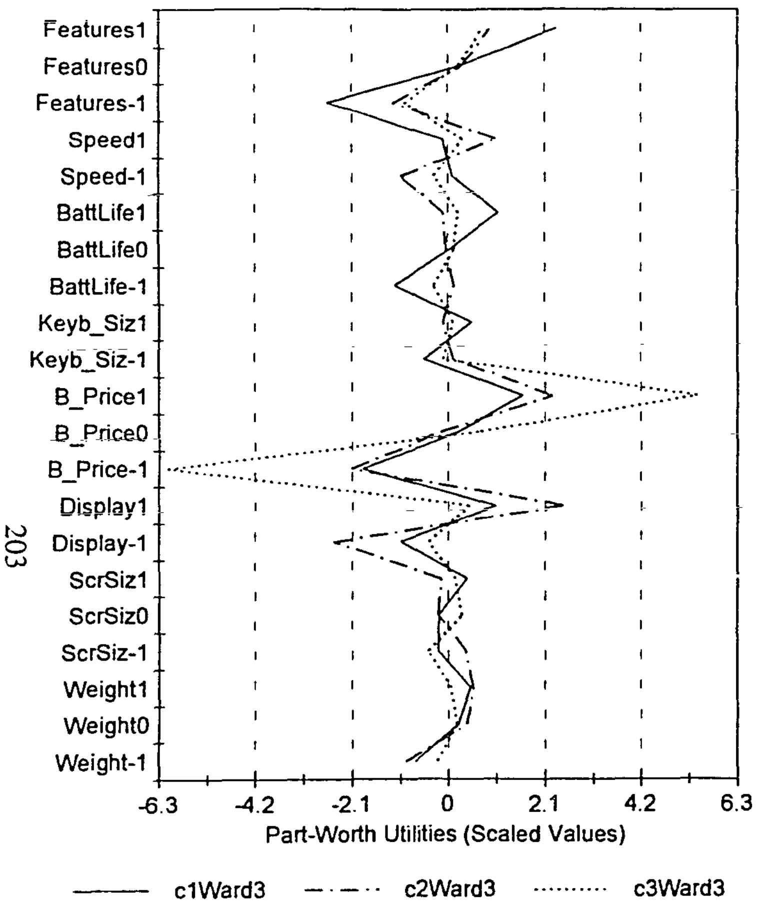


Figure 10. HIC Cluster-Profiles for 3 Clusters with Ward's Method.

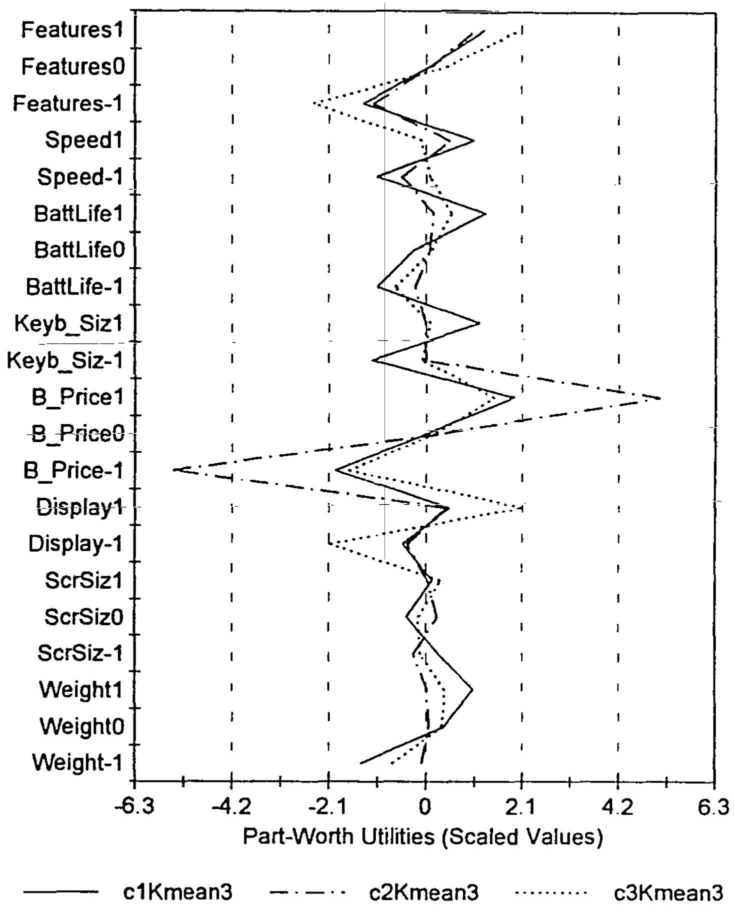


Figure 11. NHC Cluster-Profiles for 3 Clusters with Kmean Method.

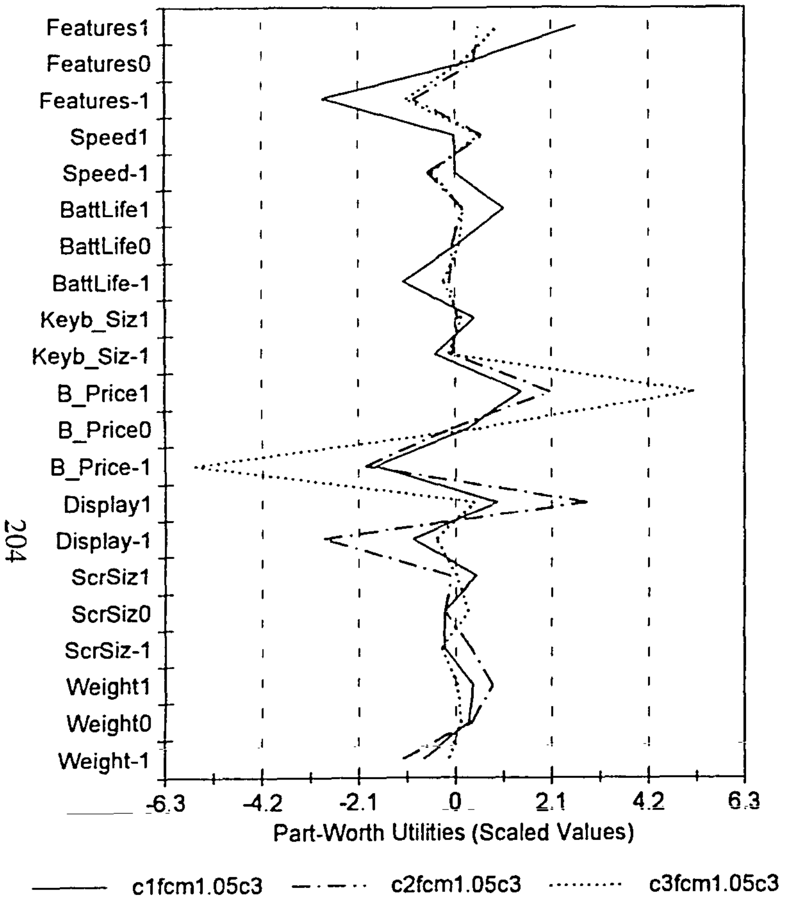


Figure 12. FUC Cluster-Profiles for 3 Clusters with m = 1.05.

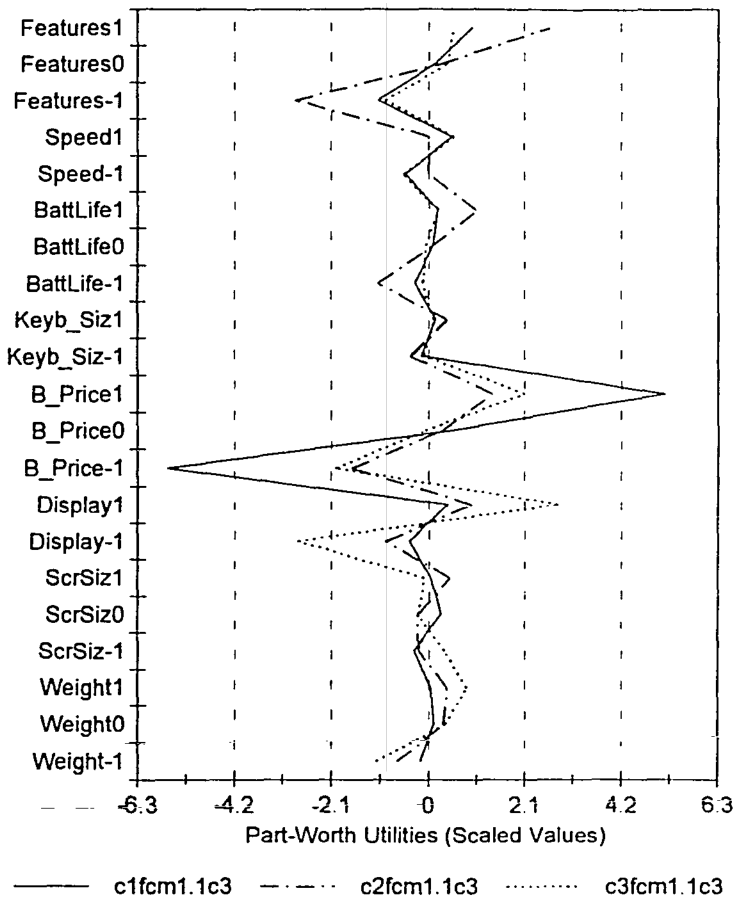


Figure 13. FUC Cluster-Profiles for 3 Clusters with m = 1.1.

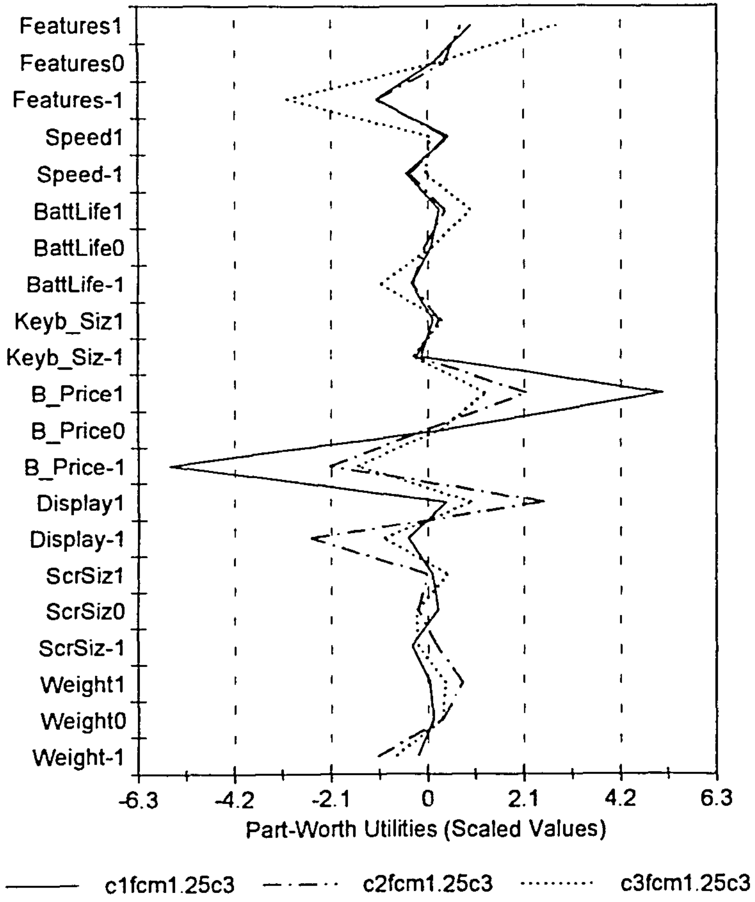


Figure 14. FUC Cluster-Profiles for 3 Clusters with m = 1.25.

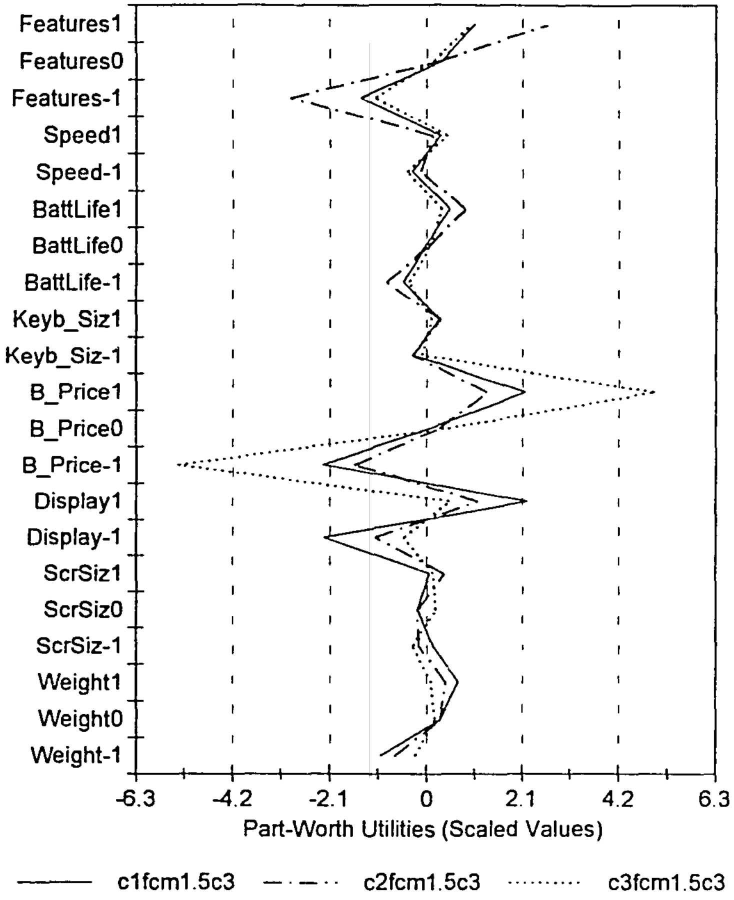


Figure 15. FUC Cluster-Profiles for 3 Clusters with m = 1.5.

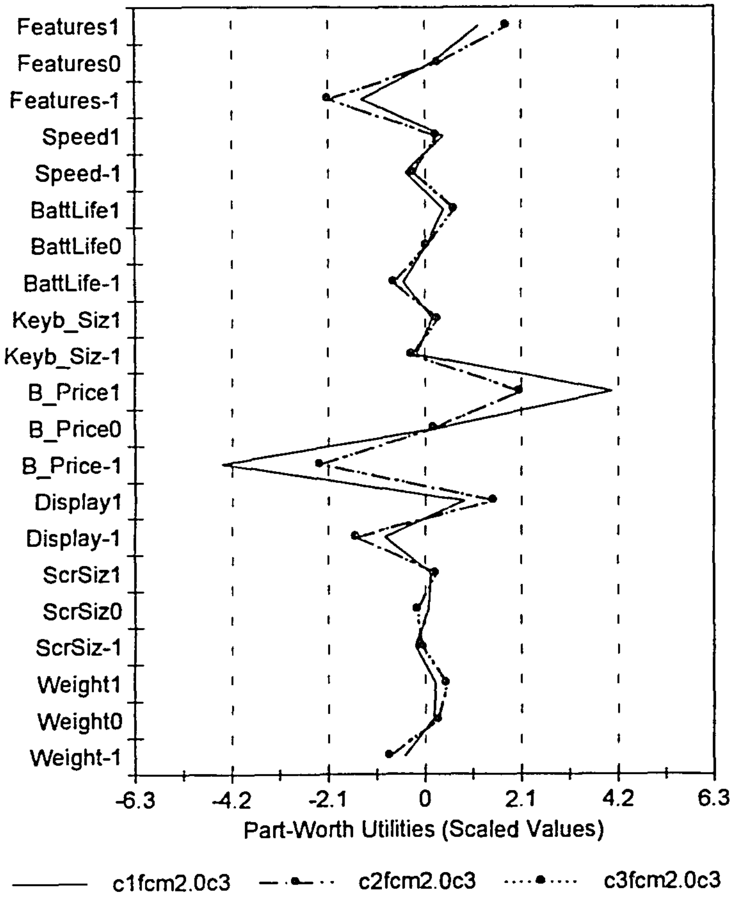


Figure 16. FUC Cluster-Profiles for 3 Clusters with m = 2.0 (2nd ≈ 3rd Cluster Center).

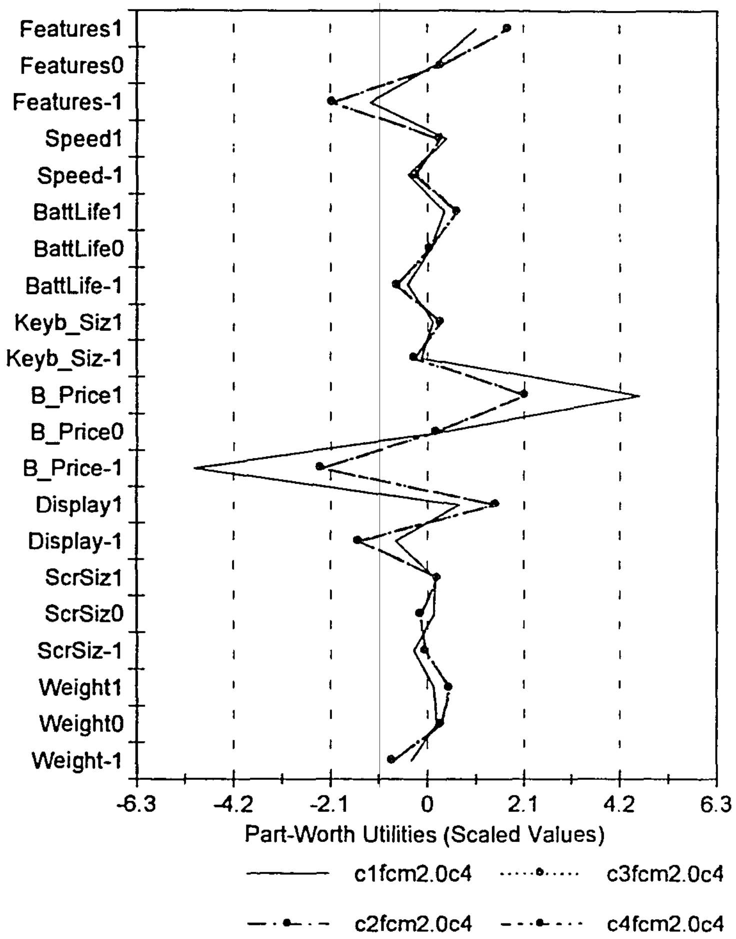


Figure 17. FUC Cluster-Profiles for 4 Clusters with m = 2.0 (2nd ≈ 3rd ≈ 4th Cluster Center).

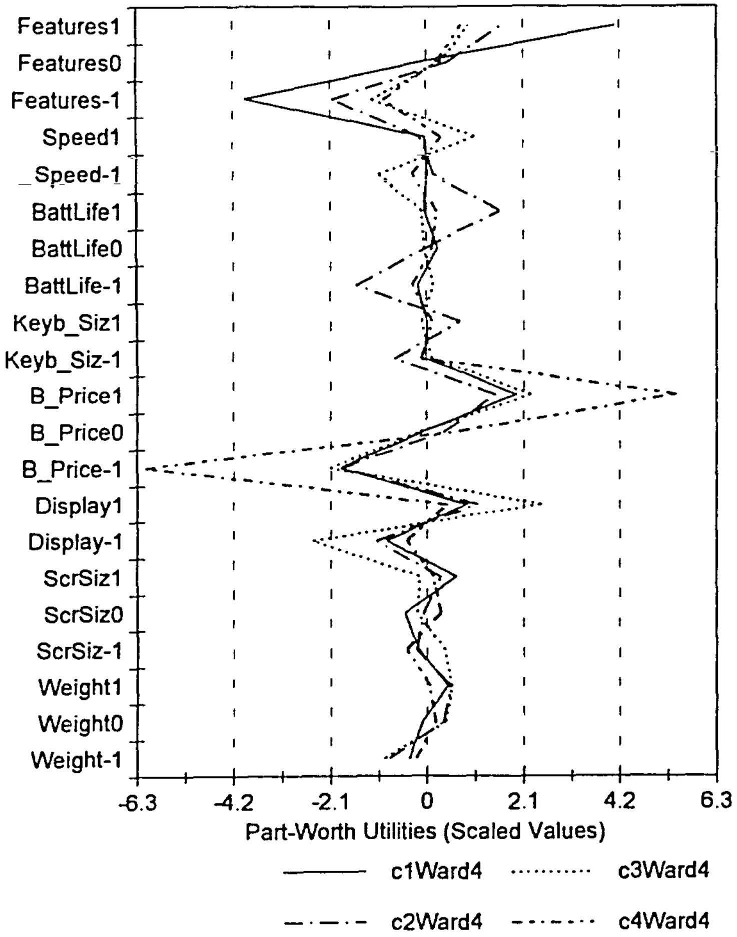


Figure 18. HIC Cluster-Profiles for 4 Clusters with Ward's Method.

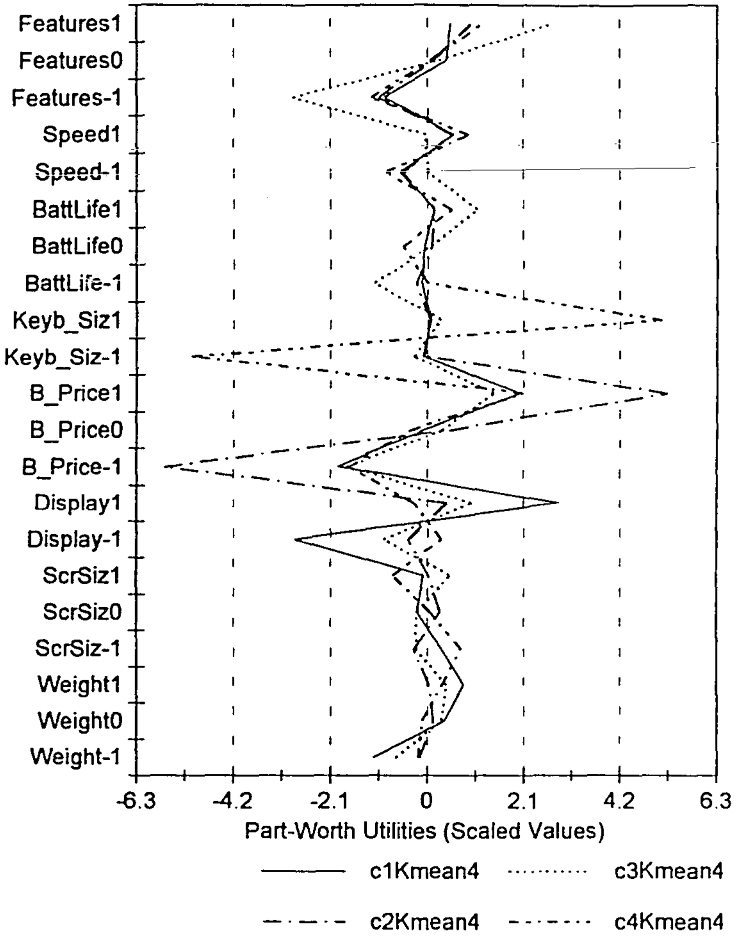


Figure 19. NHC Cluster-Profiles for 4 Clusters with Kmean Method.

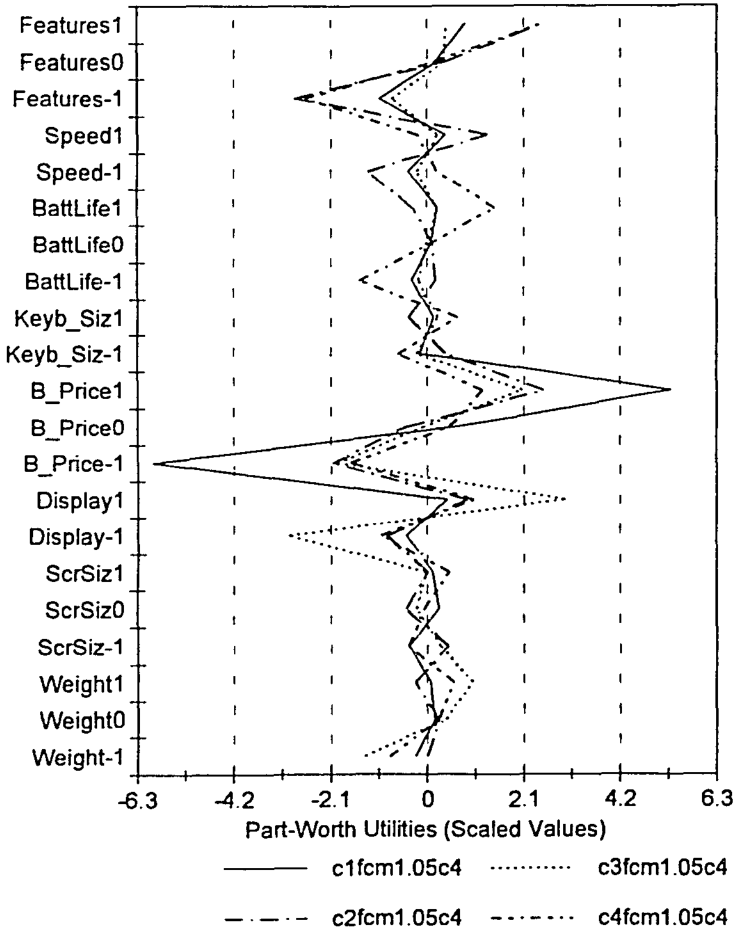


Figure 20. FUC Cluster-Profiles for 4 Clusters with m = 1.05.

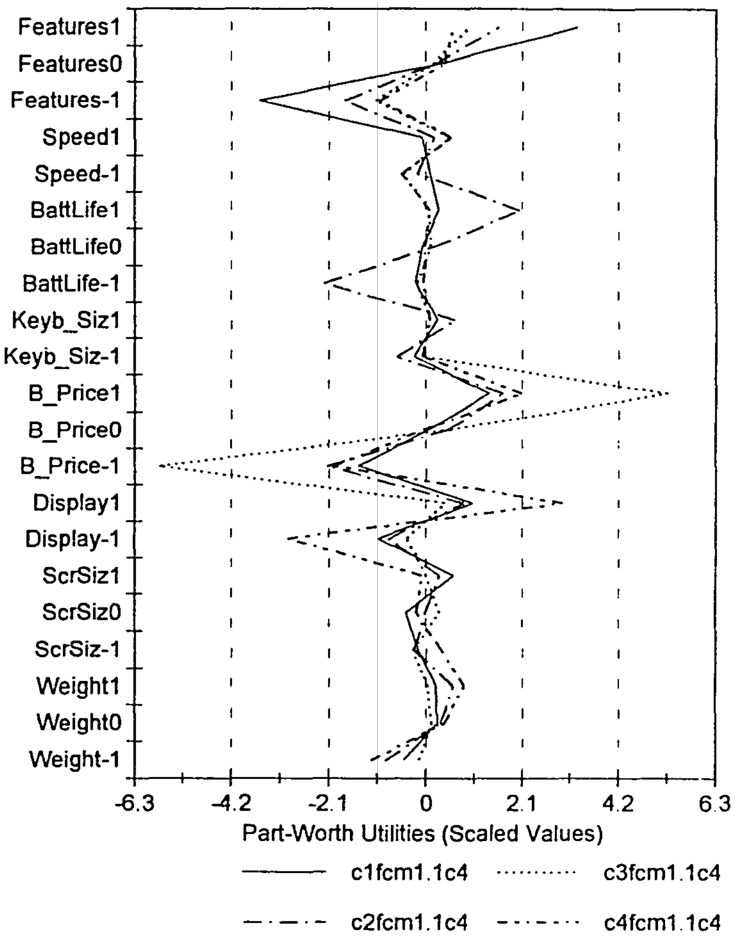


Figure 21. FUC Cluster-Profiles for 4 Clusters with m = 1.1.

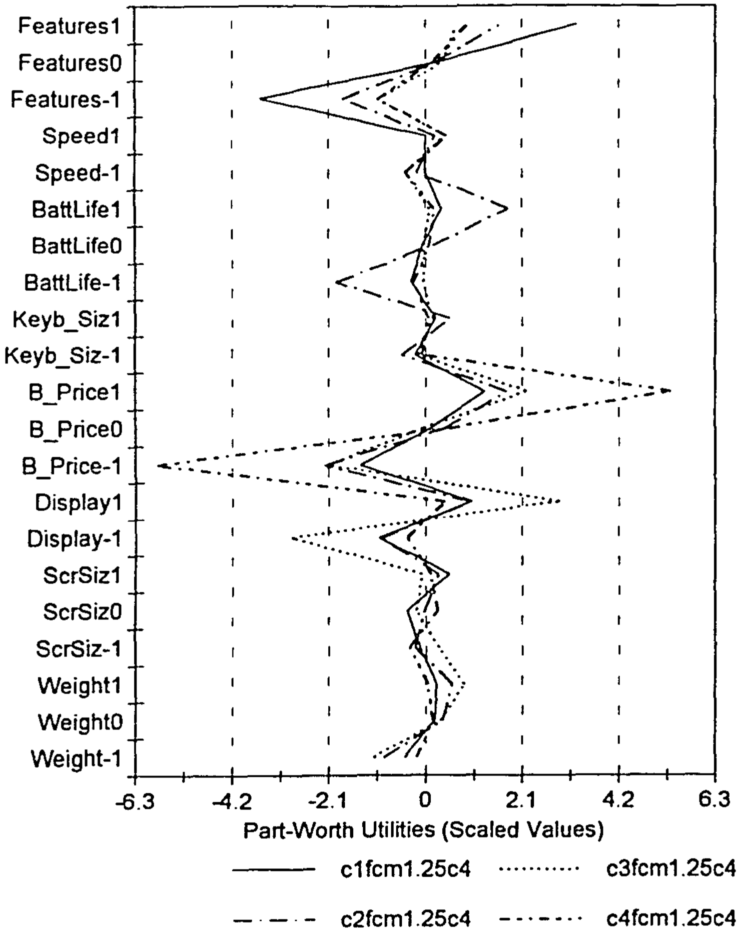


Figure 22. FUC Cluster-Profiles for 4 Clusters with m = 1.25.

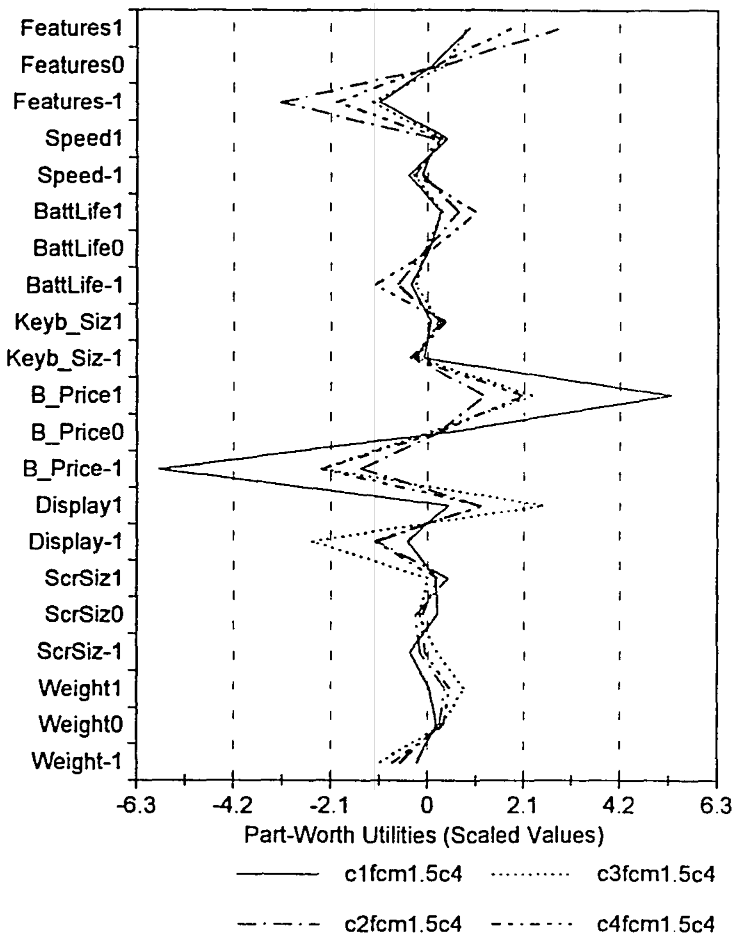


Figure 23. FUC Cluster-Profiles for 4 Clusters with m = 1.5.

TABLE XLII

COMPATIBILITY OF CLUSTERS IN PERCENTAGES OF OVERLAP

Compatibility		HIC:			NHC:			FUC:			FUC:			FUC:			FUC:			FUC:			HIC:		
of Clusters		Ward, 3 Clusters			Kmean, 3 Clusters			m=1.05, 3 Clusters			m=1.1, 3 Clusters			m=1.25, 3 Clusters			m=1.5, 3 Clusters			m=2.0, 3 Clusters			Ward, 4 Clusters		
		c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3
HIC:	c1	100.0	-	-	25.9	6.9	67.2	82.8	12.1	5.2	5.2	82.8	12.1	6.9	17.2	75.9	22.4	69.0	8.6	12.1	67.2	20.7	29.3	70.7	-
Ward, 3	c2	-	100.0	-	18.8	15.6	65.6	12.5	75.0	12:5	12.5	12.5	75.0	12.5	75.0	12.5	71.9	15.6	12.5	25.0	21.9	53.1	-	-	100.0
Clusters	c3	-	-	100.0	-	100.0	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	-	-	-
NHC:	c1	71.4	28.6	-	100.0	-	-	71.4	23.8	4.8	4.8	71.4	23.8	9.5	33.3	57.1	38.1	52.4	9.5	19.0	47.6	33.3	4.8	66.7	28.6
Kmean, 3	c2	11.1	13.9	75.0	-	100.0	-	5.6	2.8	91.7	91.7	5.6	2.8	91.7	2.8	5.6	2.8	2.8	94.4	100.0	-	-	5.6	5.6	13.9
Clusters	c3	65.0	35.0	-	-	-	100.0	58.3	41.7	-	-	58.3	41.7	-	43.3	56.7	45.0	55.0	-	3.3	60.0	36.7	23.3	41.7	35.0
FUC:	c1	92.3	7.7	-	28.8	3.8	67.3	100.0	-	-	-	100.0	-	1.9	5.8	92.3	11.5	84.6	3.8	5.8	78.8	15.4	30.8	61.5	7.7
m=1.05, 3	c2	22.6	77.4	-	16.1	3.2	80.6	-	100.0	-	-	-	100.0	-	100.0	-	96.8	3.2	-	16.1	16.1	67.7	-	22.6	77.4
Clusters	c3	8.8	11.8	79.4	2.9	97.1	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	2.9	5.9	11.8
FUC:	c1	8.8	11.8	79.4	2.9	97.1	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	2.9	5.9	11.8
m=1.1, 3	c2	92.3	7.7	-	28.8	3.8	67.3	100.0	-	-	-	100.0	-	1.9	5.8	92.3	11.5	84.6	3.8	5.8	78.8	15.4	30.8	61.5	7.7
Clusters	c3	22.6	77.4	-	16.1	3.2	80.6	-	100.0	-	-	-	100.0	-	100.0	-	96.8	3.2	-	16.1	16.1	67.7	-	22.6	77.4
FUC:	c1	11.4	11.4	77.1	5.7	94.3	-	2.9	-	97.1	97.1	2.9	-	100.0	-	-	-	-	100.0	100.0.	-	-	2.9	8.6	11.4
m=1.25, 3	c2	29.4	70.6	-	20.6	2.9	76.5	8.8	91.2	-	-	8.8	91.2	-	100.0	-	97.1	2.9	-	14.7	14.7	70.6	-	29.4	70.6
Clusters	c3	91.7	8.3	-	25.0	4.2	70.8	100.0	-	-	-	100.0	-	-	-	100.0	6.3	91.7	2.1	4.2	85.4	10.4	33.3	58.3	8.3
FUC:	c1	36.1	63.9	-	22.2	2.8	75.0	16.7	83.3	-	-	16.7	83.3	-	91.7	8.3	100.0	-	-	13.9	11.1	75.0	2.8	33.3	63.9
m=1.5, 3	c2	88.9	11.1	-	24.4	2.2	73.3	97.8	2.2	-	-	97.8	2.2	-	2.2	97.8	-	100.0	-	2.2	93.3	4.4	33.3	55.6	11.1
Clusters	c3	13.9	11.1	75.0	5.6	94.4	-	5.6	-	94.4	94.4	5.6	-	97.2	-	2.8	-	-	100.0	100.0	-	-	2.8	11.1	11.1
FUC:	c1	16.7	19.0	64.3	9.5	85.7	4.8	7.1	11.9	81.0	81.0	7.1	11.9	83.3	11.9	4.8	11.9	2.4	85.7	100.0	-	-	4.8	11.9	19.0
m=2.0, 3	c2	84.8	15.2	-	21.7	-	78.3	89.1	10.9	-	-	89.1	10.9	-	10.9	89.1	8.7	91.3	-	-	100.0	-	28.3	56.5	15.2
Clusters	c3	41.4	58.6	-	24.1	-	75.9	27.6	72.4	-	-	27.6	72.4	-	82.8	17.2	93.1	6.9	-	-	-	100.0	6.9	34.5	58.6
	c1	100.0	-	-	5.9	11.8	82.4	94.1	-	5.9	5.9	94.1	-	5.9	-	94.1	5.9	88.2	5.9	11.8	76.5	11.8	100.0	-	-
HIC:	c2	100.0	-	-	34.1	4.9	61.0	78.0	17.1	4.9	4.9	78.0	17.1	7.3	24.4	68.3	29.3	61.0	9.8	12.2	63.4	24.4	-	100.0	-
Ward, 4	c3	-	100.0	-	18.8	15.6	65.6	12.5	75.0	12.5	12.5	12.5	75.0	12.5	75.0	12.5	71.9	15.6	12.5	25.0	21.9	53.1	-	-	100.0
Clusters	c4	-	-	100.0	-	100.0	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	-	-	-
	c1	22.6	77.4	-	16.1	3.2	80.6	-	100.0	-	-	-	100.0	-	100.0	-	96.8	3.2	-	16.1	16.1	67.7	-	22.6	77.4
NHC:	c2	6.1	12.1	81.8	-	100.0	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	3.0	3.0	12.1
Kmean, 4	c3	92.2	7.8	-	27.5	3.9	68.6	100.0	-	-	-	100.0	-	2.0	5.9	92.2	11.8	84.3	3.9	5.9	78.4	15.7	31.4	60.8	7.8
Clusters	c4	100.0	-	-	100.0	-	-	50.0	-	50.0	50.0	50.0	-	50.0	-	50.0	-	50.0	50.0	50.0	50.0	-	-	100.0	-
	c1	6.7	3.3	90.0	-	100.0	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	3.3	3.3	3.3
FUC:	c2	45.0	55.0	-	25.0	25.0	50.0	60.0	20.0	20.0	20.0	60.0	20.0	20.0	20.0	60.0	15.0	65.0	20.0	30.0	50.0	20.0	35.0	10.0	55.0
m=1.05, 4	c3	25.9	74.1	-	14.8	-	85.2	-	100.0	-	-	-	100.0	-	100.0	-	100.0	-	-	14.8	14.8	70.4	-	25.9	74.1
Clusters	c4	100.0	-	-	30.0	2.5	67.5	100.0	-	-	-	100.0	-	2.5	7.5	90.0	15.0	80.0	5.0	5.0	80.0	15.0	22.5	77.5	-
	c1	89.7	10.3	-	17.2	3.4	79.3	100.0	-	-	-	100.0	-	-	-	100.0	3.4	96.6	-	3.4	86.2	10.3	55.2	34.5	10.3
FUC:	c2	96.2	3.8	-	46.2	7.7	46.2	88.5	3.8	7.7	7.7	88.5	3.8	11.5	15.4	73.1	23.1	61.5	15.4	19.2	61.5	19.2	-	96.2	3.8
m=1.1, 4	c3	3.1	12.5	84.4	-	100.0	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	3.1	-	12.5
Clusters	c4	20.0	80.0	-	13.3	3.3	83.3	-	100.0	-	-	-	100.0	-	100.0	-	96.7	3.3	-	13.3	16.7	70.0	-	20.0	80.0
	c1	86.7	13.3	-	16.7	3.3	80.0	96.7	3.3	-	-	96.7	3.3	-	3.3	96.7	3.3	96.7	-	3.3	86.7	10.0	53.3	33.3	13.3
FUC:	c2	96.2	3.8	-	46.2	7.7	46.2	88.5	3.8	7.7	7.7	88.5	3.8	11.5	15.4	73.1	23.1	61.5	15.4	19.2	61.5	19.2	-	96.2	3.8
m=1.25, 4	c3	20.7	79.3	-	13.8	3.4	82.8	-	100.0	-	-	-	100.0	-	100.0	-	100.0	-	-	13.8	13.8	72.4	-	20.7	79.3
Clusters	c4	3.1	12.5	84.4	-	100.0	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	3.1	-	12.5
	c1	3.2	9.7	87.1	-	100.0	-	-	-	100.0	100.0	-	-	100.0	-	-	-	-	100.0	100.0	-	-	3.2	-	9.7
FUC:	c2	86.2	13.8	-	13.8	3.4	82.8	96.6	3.4	-	-	96.6	3.4	-	3.4	96.6	-	100.0	-	3.4	96.6	-	48.3	37.9	13.8
m=1.5, 4	c3	20.7	79.3	-	10.3	6.9	82.8	-	96.6	3.4	3.4	-	96.6	3.4	96.6	-	96.6	-	3.4	17.2	13.8	69.0	-	20.7	79.3
Clusters	c4	92.9	7.1	-	50.0	7.1	42.9	85.7	7.1	7.1	7.1	85.7	7.1	10.7	17.9	71.4	28.6	57.1	14.3	17.9	50.0	32.1	7.1	85.7	7.1
	c1	11.1	13.9	75.0	2.8	97.2	-	2.8	5.6	91.7	91.7	2.8	5.6	91.7	5.6	2.8	5.6	-	94.4	100.0	-	-	2.8	8.3	13.9
FUC:	c2	83.3	16.7	-	24.1	1.9	74.1	85.2	14.8	-	-	85.2	14.8	-	14.8	85.2	16.7	83.3	-	1.9	85.2	13.0	27.8	55.6	16.7
m=2.0, 4	c3	33.3	66.7	-	25.9	-	74.1	18.5	77.8	3.7	3.7	18.5	77.8	7.4	88.9	3.7	92.6	-	7.4	18.5	-	81.5	3.7	29.6	66.7
Clusters	c4	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

n/a: Though cluster c4f2.0c4 is theoretically possible, it did not materialize with this empirical data set. Determination of cluster membership with the maximum membership in any one cluster "re-"classified most of responses in column 4 of partition matrix U into members of cluster 3.

Interpretation of percentages of cluster compatibility is as follows:

Each cluster cell which corresponds to a two-way table of clustering methods lists the percentage of respondents in the row cluster who are members in the column cluster.

	HIC:				NHC:				FUC:				FUC:				FUC:				FUC:				FUC:			
usters	Ward, 4 Clusters				Kmean, 4 Clusters				m=1.05, 4 Clusters				m=1.1, 4 Clusters				m=1.25, 4 Clusters				m=1.5, 4 Clusters				m=2.0, 4 Clusters			
c3	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4
20.7	29.3	70.7	-	-	12.1	3.4	81.0	3.4	3.4	15.5	12.1	69.0	44.8	43.1	1.7	10.3	44.8	43.1	10.3	1.7	1.7	43.1	10.3	44.8	6.9	77.6	15.5	n/a
53.1	-	-	100.0	-	75.0	12.5	12.5	-	3.1	34.4	62.5	-	9.4	3.1	12.5	75.0	12.5	3.1	71.9	12.5	9.4	12.5	71.9	6.3	15.6	28.1	56.3	n/a
-	-	-	-	100.0	-	100.0	-	-	100.0	-	-	-	-	-	100.0	-	-	-	-	100.0	100.0	-	-	-	100.0	-	-	n/a
33.3	4.8	66.7	28.6	-	23.8	-	66.7	9.5	-	23.8	19.0	57.1	23.8	57.1	-	19.0	23.8	57.1	19.0	-	-	19.0	14.3	66.7	4.8	61.9	33.3	n/a
-	5.6	5.6	13.9	75.0	2.8	91.7	5.6	-	83.3	13.9	-	2.8	2.8	5.6	88.9	2.8	2.8	5.6	2.8	88.9	86.1	2.8	5.6	5.6	97.2	2.8	-	n/a
36.7	23.3	41.7	35.0	-	41.7	-	58.3	-	-	16.7	38.3	45.0	38.3	20.0	-	41.7	40.0	20.0	40.0	-	-	40.0	40.0	20.0	-	66.7	33.3	n/a
15.4	30.8	61.5	7.7	-	-	-	98.1	1.9	-	23.1	-	76.9	55.8	44.2	-	-	55.8	44.2	-	-	-	53.8	-	46.2	1.9	88.5	9.6	n/a
67.7	-	22.6	77.4	-	100.0	-	-	-	-	12.9	87.1	-	-	3.2	-	96.8	3.2	3.2	93.5	-	-	3.2	90.3	6.5	6.5	25.8	67.7	n/a
-	2.9	5.9	11.8	79.4	-	97.1	-	2.9	88.2	11.8	-	-	-	5.9	94.1	-	-	5.9	-	94.1	91.2	-	2.9	5.9	97.1	-	2.9	n/a
-	2.9	5.9	11.8	79.4	-	97.1	-	2.9	88.2	11.8	-	-	-	5.9	94.1	-	-	5.9	-	94.1	91.2	-	2.9	5.9	97.1	-	2.9	n/a
15.4	30.8	61.5	7.7	-	-	-	98.1	1.9	-	23.1	-	76.9	55.8	44.2	-	-	55.8	44.2	-	-	-	53.8	-	46.2	1.9	88.5	9.6	n/a
67.7	-	22.6	77.4	-	100.0	-	-	-	-	12.9	87.1	-	-	3.2	-	96.8	3.2	3.2	93.5	-	-	3.2	90.3	6.5	6.5	25.8	67.7	n/a
-	2.9	8.6	11.4	77.1	-	94.3	2.9	2.9	85.7	11.4	-	2.9	-	8.6	91.4	-	-	8.6	-	91.4	88.6	-	2.9	8.6	94.3	-	5.7	n/a
70.6	-	29.4	70.6	-	91.2	-	8.8	-	-	11.8	79.4	8.8	-	11.8	-	88.2	2.9	11.8	85.3	-	-	2.9	82.4	14.7	5.9	23.5	70.6	n/a
10.4	33.3	58.3	8.3	-	-	-	97.9	2.1	-	25.0	-	75.0	60.4	39.6	-	-	60.4	39.6	-	-	-	58.3	-	41.7	2.1	95.8	2.1	n/a
75.0	2.8	33.3	63.9	-	83.3	-	16.7	-	-	8.3	75.0	16.7	2.8	16.7	-	80.6	2.8	16.7	80.6	-	-	-	77.8	22.2	5.6	25.0	69.4	n/a
4.4	33.3	55.6	11.1	-	2.2	-	95.6	2.2	-	28.9	-	71.1	62.2	35.6	-	2.2	64.4	35.6	-	-	-	64.4	-	35.6	-	100.0	-	n/a
-	2.8	11.1	11.1	75.0	-	91.7	5.6	2.8	83.3	11.1	-	5.6	-	11.1	88.9	-	-	11.1	-	88.9	86.1	-	2.8	11.1	94.4	-	5.6	n/a
-	4.8	11.9	19.0	64.3	11.9	78.6	7.1	2.4	71.4	14.3	9.5	4.8	2.4	11.9	76.2	9.5	2.4	11.9	9.5	76.2	73.8	2.4	11.9	11.9	85.7	2.4	11.9	n/a
-	28.3	56.5	15.2	-	10.9	-	87.0	2.2	-	21.7	8.7	69.6	54.3	34.8	-	10.9	56.5	34.8	8.7	-	-	60.9	8.7	30.4	-	100.0	-	n/a
100.0	6.9	34.5	58.6	-	72.4	-	27.6	-	-	13.8	65.5	20.7	10.3	17.2	-	72.4	10.3	17.2	72.4	-	-	-	69.0	31.0	-	24.1	75.9	n/a
11.8	100.0	-	-	-	-	5.9	94.1	-	5.9	41.2	-	52.9	94.1	-	5.9	-	94.1	-	-	5.9	5.9	82.4	-	11.8	5.9	88.2	5.9	n/a
24.4	-	100.0	-	-	17.1	2.4	75.6	4.9	2.4	4.9	17.1	75.6	24.4	61.0	-	14.6	24.4	61.0	14.6	-	-	26.8	14.6	58.5	7.3	73.2	19.5	n/a
53.1	-	-	100.0	-	75.0	12.5	12.5	-	3.1	34.4	62.5	-	9.4	3.1	12.5	75.0	12.5	3.1	71.9	12.5	9.4	12.5	71.9	6.3	15.6	28.1	56.3	n/a
-	-	-	-	100.0	-	100.0	-	-	100.0	-	-	-	-	-	100.0	-	-	-	-	100.0	100.0	-	-	-	100.0	-	-	n/a
67.7	-	22.6	77.4	-	100.0	-	-	-	-	12.9	87.1	-	-	3.2	-	96.8	3.2	3.2	93.5	-	-	3.2	90.3	6.5	6.5	25.8	67.7	n/a
-	3.0	3.0	12.1	81.8	-	100.0	-	-	90.9	9.1	-	-	-	3.0	97.0	-	-	3.0	-	97.0	93.9	-	3.0	3.0	100.0	-	-	n/a
15.7	31.4	60.8	7.8	-	-	-	100.0	-	-	23.5	-	76.5	56.9	43.1	-	-	56.9	43.1	-	-	-	54.9	-	45.1	2.0	88.2	9.8	n/a
-	-	100.0	-	-	-	-	-	100.0	-	50.0	-	50.0	-	100.0	-	-	-	100.0	-	-	-	-	-	100.0	-	50.0	50.0	n/a
-	3.3	3.3	3.3	90.0	-	100.0	-	-	100.0	-	-	-	-	3.3	96.7	-	-	3.3	-	96.7	96.7	-	-	3.3	100.0	-	-	n/a
20.0	35.0	10.0	55.0	-	20.0	15.0	60.0	5.0	-	100.0	-	-	55.0	10.0	15.0	20.0	60.0	10.0	15.0	15.0	10.0	50.0	15.0	25.0	20.0	70.0	10.0	n/a
70.4	-	25.9	74.1	-	100.0	-	-	-	-	-	100.0	-	-	3.7	-	96.3	-	3.7	96.3	-	-	-	96.3	3.7	3.7	22.2	74.1	n/a
15.0	22.5	77.5	-	-	-	-	97.5	2.5	-	-	-	100.0	45.0	55.0	-	-	45.0	55.0	-	-	-	47.5	-	52.5	2.5	85.0	12.5	n/a
10.3	55.2	34.5	10.3	-	-	-	100.0	-	-	37.9	-	62.1	100.0	-	-	-	100.0	-	-	-	-	86.2	-	13.8	-	96.6	3.4	n/a
10.2	-	96.2	3.8	-	3.8	3.8	84.6	7.7	3.8	7.7	3.8	84.6	-	100.0	-	-	-	100.0	-	-	-	11.5	-	88.5	11.5	69.2	19.2	n/a
-	3.1	-	12.5	84.4	-	100.0	-	-	90.6	9.4	-	-	-	-	100.0	-	-	-	-	100.0	96.9	-	3.1	-	100.0	-	-	n/a
70.0	-	20.0	80.0	-	100.0	-	-	-	-	13.3	86.7	-	-	-	-	100.0	3.3	-	96.7	-	-	3.3	93.3	3.3	3.3	26.7	70.0	n/a
10.0	53.3	33.3	13.3	-	3.3	-	96.7	-	-	40.0	-	60.0	96.7	-	-	3.3	100.0	-	-	-	-	86.7	-	13.3	-	96.7	3.3	n/a
19.2	-	96.2	3.8	-	3.8	3.8	84.6	7.7	3.8	7.7	3.8	84.6	-	100.0	-	-	-	100.0	-	-	-	11.5	-	88.5	11.5	69.2	19.2	n/a
72.4	-	20.7	79.3	-	100.0	-	-	-	-	10.3	89.7	-	-	-	-	100.0	-	-	100.0	-	-	-	96.6	3.4	3.4	24.1	72.4	n/a
-	3.1	-	12.5	84.4	-	100.0	-	-	90.6	9.4	-	-	-	-	100.0	-	-	-	-	100.0	96.9	-	3.1	-	100.0	-	-	n/a
-	3.2	-	9.7	87.1	-	100.0	-	-	93.5	6.5	-	-	-	-	100.0	-	-	-	-	100.0	100.0	-	-	-	100.0	-	-	n/a
-	48.3	37.9	13.8	-	3.4	-	96.6	-	-	34.5	-	65.5	86.2	10.3	-	3.4	89.7	10.3	-	-	-	100.0	-	-	-	100.0	-	n/a
69.0	-	20.7	79.3	-	96.6	3.4	-	-	-	10.3	89.7	-	-	-	3.4	96.6	-	-	96.6	3.4	-	-	100.0	-	6.9	20.7	72.4	n/a
32.1	7.1	85.7	7.1	-	7.1	3.6	82.1	7.1	3.6	17.9	3.6	75.0	14.3	82.1	-	3.6	14.3	82.1	3.6	-	-	-	-	100.0	10.7	67.9	21.4	n/a
-	2.8	8.3	13.9	75.0	5.6	91.7	2.8	-	83.3	11.1	2.8	2.8	-	8.3	88.9	2.8	-	8.3	2.8	88.9	86.1	-	5.6	8.3	100.0	-	-	n/a
13.0	27.8	55.6	16.7	-	14.8	-	83.3	1.9	-	25.9	11.1	63.0	51.9	33.3	-	14.8	53.7	33.3	13.0	-	-	53.7	11.1	35.2	-	100.0	-	n/a
81.5	3.7	29.6	66.7	-	77.8	-	18.5	3.7	-	7.4	74.1	18.5	3.7	18.5	-	77.8	3.7	18.5	77.8	-	-	-	77.8	22.2	-	-	100.0	n/a
n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

TABLE XLIII

COMPATIBILITY OF CLUSTERS IN ABSOLUTE NUMBERS

Compatibility of Clusters		HIC: Ward, 3 Clusters			NHC: Kmean, 3 Clusters			FUC: m=1.05, 3 Clusters			FUC: m=1.1, 3 Clusters			FUC: m=1.25, 3 Clusters			FUC: m=1.5, 3 Clusters			FUC: m=2.0, 3 Clusters			HIC: Ward, 4 Clusters			
		c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c4
HIC:	c1	**58**	-	-	15	4	**39**	**48**	7	3	3	**48**	7	4	10	**44**	13	**40**	5	7	**39**	12	17	**41**	-	-
Ward, 3	c2	-	**32**	-	6	5	**21**	4	**24**	4	4	4	**24**	4	**24**	4	**23**	5	4	8	7	**17**	-	-	**32**	-
Clusters	c3	-	-	**27**	-	**27**	-	-	-	**27**	**27**	-	-	**27**	-	-	-	-	**27**	**27**	-	-	-	-	-	**27**
NHC:	c1	**15**	6	-	**21**	-	-	**15**	5	1	1	**15**	5	2	7	**12**	8	**11**	2	4	**10**	7	1	**14**	6	-
Kmean, 3	c2	4	5	**27**	-	**36**	-	2	1	**33**	**33**	2	1	**33**	1	2	1	1	**34**	**36**	-	-	2	2	5	**27**
Clusters	c3	**39**	21	-	-	-	**60**	**35**	25	-	-	**35**	25	-	26	**34**	27	**33**	-	2	**36**	22	14	**25**	21	-
FUC:	c1	**48**	4	-	15	2	**35**	**52**	-	-	-	**52**	-	1	3	**48**	6	**44**	2	3	**41**	8	16	**32**	4	-
m=1.05, 3	c2	7	**24**	-	5	1	**25**	-	**31**	-	-	-	**31**	-	**31**	-	**30**	1	-	5	5	**21**	-	7	**24**	-
Clusters	c3	3	4	**27**	1	**33**	-	-	-	**34**	**34**	-	-	**34**	-	-	-	-	**34**	**34**	-	-	1	2	4	**27**
FUC:	c1	3	4	**27**	1	**33**	-	-	-	**34**	**34**	-	-	**34**	-	-	-	-	**34**	**34**	-	-	1	2	4	**27**
m=1.1, 3	c2	**48**	4	-	15	2	**35**	**52**	-	-	-	**52**	-	1	3	**48**	6	**44**	2	3	**41**	8	16	**32**	4	-
Clusters	c3	7	**24**	-	5	1	**25**	-	**31**	-	-	-	**31**	-	**31**	-	**30**	1	-	5	5	**21**	-	7	**24**	-
FUC:	c1	4	4	**27**	2	**33**	-	1	-	**34**	**34**	1	-	**35**	-	-	-	-	**35**	**35**	-	-	1	3	4	**27**
m=1.25, 3	c2	10	**24**	-	7	1	**26**	3	**31**	-	-	3	**31**	-	**34**	-	**33**	1	-	5	5	**24**	-	10	**24**	-
Clusters	c3	**44**	4	-	12	2	**34**	**48**	-	-	-	**48**	-	-	-	**48**	3	**44**	1	2	**41**	5	16	**28**	4	-
FUC:	c1	13	**23**	-	8	1	**27**	6	**30**	-	-	6	**30**	-	**33**	3	**36**	-	-	5	4	**27**	1	12	**23**	-
m=1.5, 3	c2	**40**	5	-	11	1	**33**	**44**	1	-	-	**44**	1	-	1	**44**	-	**45**	-	1	**42**	2	15	**25**	5	-
Clusters	c3	5	4	**27**	2	**34**	-	2	-	**34**	**34**	2	-	**35**	-	1	-	-	**36**	**36**	-	-	1	4	4	**27**
FUC:	c1	7	8	**27**	4	**36**	2	3	5	**34**	**34**	3	5	**35**	5	2	5	1	**36**	**42**	-	-	2	5	8	**27**
m=2.0, 3	c2	**39**	7	-	10	-	**36**	**41**	5	-	-	**41**	5	-	5	**41**	4	**42**	-	-	**46**	-	13	**26**	7	-
Clusters	c3	12	**17**	-	7	-	**22**	8	**21**	-	-	8	**21**	-	**24**	5	**27**	2	-	-	-	**29**	2	10	**17**	-
	c1	**17**	-	-	1	2	**14**	**16**	-	1	1	**16**	-	1	-	**16**	1	**15**	1	2	**13**	2	**17**	-	-	-
HIC:	c2	**41**	-	-	14	2	**25**	**32**	7	2	2	**32**	7	3	10	**28**	12	**25**	4	5	**26**	10	-	**41**	-	-
Ward, 4	c3	-	**32**	-	6	5	**21**	4	**24**	4	4	4	**24**	4	**24**	4	**23**	5	4	8	7	**17**	-	-	**32**	-
Clusters	c4	-	-	**27**	-	**27**	-	-	-	**27**	**27**	-	-	**27**	-	-	-	-	**27**	**27**	-	-	-	-	-	**27**
	c1	7	**24**	-	5	1	**25**	-	**31**	-	-	-	**31**	-	**31**	-	**30**	1	-	5	5	**21**	-	7	**24**	-
NHC:	c2	2	4	**27**	-	**33**	-	-	-	**33**	**33**	-	-	**33**	-	-	-	-	**33**	**33**	-	-	1	1	4	**27**
Kmean, 4	c3	**47**	4	-	14	2	**35**	**51**	-	-	-	**51**	-	1	3	**47**	6	**43**	2	3	**40**	8	16	**31**	4	-
Clusters	c4	**2**	-	-	**2**	-	-	**1**	-	**1**	**1**	**1**	-	**1**	-	**1**	-	**1**	**1**	**1**	**1**	-	-	**2**	-	-
	c1	2	1	**27**	-	**30**	-	-	-	**30**	**30**	-	-	**30**	-	-	-	-	**30**	**30**	-	-	1	1	1	**27**
FUC:	c2	9	**11**	-	5	5	**10**	**12**	4	4	4	**12**	4	4	4	**12**	3	**13**	4	6	**10**	4	7	2	**11**	-
m=1.05, 4	c3	7	**20**	-	4	-	**23**	-	**27**	-	-	-	**27**	-	**27**	-	**27**	-	-	4	4	**19**	-	7	**20**	-
Clusters	c4	**40**	-	-	12	1	**27**	**40**	-	-	-	**40**	-	1	3	**36**	6	**32**	2	2	**32**	6	9	**31**	-	-
	c1	**26**	3	-	5	1	**23**	**29**	-	-	-	**29**	-	-	-	**29**	1	**28**	-	1	**25**	3	**16**	10	3	-
FUC:	c2	**25**	1	-	**12**	2	**12**	**23**	1	2	2	**23**	1	3	4	**19**	6	**16**	4	5	**16**	5	-	**25**	1	-
m=1.1, 4	c3	1	4	**27**	-	**32**	-	-	-	**32**	**32**	-	-	**32**	-	-	-	-	**32**	**32**	-	-	1	-	4	**27**
Clusters	c4	6	**24**	-	4	1	**25**	-	**30**	-	-	-	**30**	-	**30**	-	**29**	1	-	4	5	**21**	-	6	**24**	-
	c1	**26**	4	-	5	1	**24**	**29**	1	-	-	**29**	1	-	1	**29**	1	**29**	-	1	**26**	3	**16**	10	4	-
FUC:	c2	**25**	1	-	**12**	2	**12**	**23**	1	2	2	**23**	1	3	4	**19**	6	**16**	4	5	**16**	5	-	**25**	1	-
m=1.25, 4	c3	6	**23**	-	4	1	**24**	-	**29**	-	-	-	**29**	-	**29**	-	**29**	-	-	4	4	**21**	-	6	**23**	-
Clusters	c4	1	4	**27**	-	**32**	-	-	-	**32**	**32**	-	-	**32**	-	-	-	-	**32**	**32**	-	-	1	-	4	**27**
	c1	1	3	**27**	-	**31**	-	-	-	**31**	**31**	-	-	**31**	-	-	-	-	**31**	**31**	-	-	1	-	3	**27**
FUC:	c2	**25**	4	-	4	1	**24**	**28**	1	-	-	**28**	1	-	1	**28**	-	**29**	-	1	**28**	-	**14**	11	4	-
m=1.5, 4	c3	6	**23**	-	3	2	**24**	-	**28**	1	1	-	**28**	1	**28**	-	**28**	-	1	5	4	**20**	-	6	**23**	-
Clusters	c4	**26**	2	-	**14**	2	12	**24**	2	2	2	**24**	2	3	5	**20**	8	**16**	4	5	**14**	9	2	**24**	2	-
	c1	4	5	**27**	1	**35**	-	1	2	**33**	**33**	1	2	**33**	2	1	2	-	**34**	**36**	-	-	1	3	5	**27**
FUC:	c2	**45**	9	-	13	1	**40**	**46**	8	-	-	**46**	8	-	8	**46**	9	**45**	-	1	**46**	7	15	**30**	9	-
m=2.0, 4	c3	9	**18**	-	7	-	**20**	5	**21**	1	1	5	**21**	2	**24**	1	**25**	-	2	5	-	**22**	1	8	**18**	-
Clusters	c4	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Interpretation of absolute numbers of cluster compatibility is as follows:

Each cluster cell which corresponds to a two-way table of clustering methods lists the absolute number of respondents in the row cluster who are members in the column cluster.

FUC: 3 Clusters		HIC: Ward, 4 Clusters				NHC: Kmean, 4 Clusters				FUC: m=1.05, 4 Clusters				FUC: m=1.1, 4 Clusters				FUC: m=1.25, 4 Clusters				FUC: m=1.5, 4 Clusters				FUC: m=2.0, 4 Clusters			
c2	c3	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4
39	12	17	41	-	-	7	2	47	2	2	9	7	40	26	25	1	6	26	25	6	1	1	25	6	26	4	45	9	n/a
7	17	-	-	32	-	24	4	4	-	1	11	20	-	3	1	4	24	4	1	23	4	3	4	23	2	5	9	18	n/a
-	-	-	-	-	27	-	27	-	-	27	-	-	-	-	-	27	-	-	-	-	27	27	-	-	-	27	-	-	n/a
10	7	1	14	6	-	5	-	14	2	-	5	4	12	5	12	-	4	5	12	4	-	-	4	3	14	1	13	7	n/a
-	-	2	2	5	27	1	33	2	-	30	5	-	1	1	2	32	1	1	2	1	32	31	1	2	2	35	1	-	n/a
36	22	14	28	21	-	25	-	35	-	-	10	23	27	23	12	-	25	24	12	24	-	-	24	24	12	-	40	20	n/a
41	8	16	32	4	-	-	-	51	1	-	12	-	40	29	23	-	-	29	23	-	-	-	28	-	24	1	46	5	n/a
5	21	-	7	24	-	31	-	-	-	-	4	27	-	-	1	-	30	1	1	29	-	-	1	28	2	2	8	21	n/a
-	-	1	2	4	27	-	33	-	1	30	4	-	-	-	2	32	-	-	2	-	32	31	-	1	2	33	-	1	n/a
-	-	1	2	4	27	-	33	-	1	30	4	-	-	-	2	32	-	-	2	-	32	31	-	1	2	33	-	1	n/a
41	8	16	32	4	-	-	-	51	1	-	12	-	40	29	23	-	-	29	23	-	-	-	28	-	24	1	46	5	n/a
5	21	-	7	24	-	31	-	-	-	-	4	27	-	-	1	-	30	1	1	29	-	-	1	28	2	2	8	21	n/a
-	-	1	3	4	27	-	33	1	1	30	4	-	1	-	3	32	-	-	3	-	32	31	-	1	3	33	-	2	n/a
5	24	-	10	24	-	31	-	3	-	-	4	27	3	-	4	-	30	1	4	29	-	-	1	28	5	2	8	24	n/a
41	5	16	28	4	-	-	-	47	1	-	12	-	36	29	19	-	-	29	19	-	-	-	28	-	20	1	45	1	n/a
4	27	1	12	23	-	30	-	6	-	-	3	27	6	1	6	-	29	1	6	29	-	-	-	28	8	2	9	25	n/a
42	2	15	25	5	-	1	-	43	1	-	13	-	32	28	16	-	1	29	16	-	-	-	29	-	16	-	45	-	n/a
-	-	1	4	4	27	-	33	2	1	30	4	-	2	-	4	32	-	-	4	-	32	31	-	1	4	34	-	2	n/a
-	-	2	5	8	27	5	33	3	1	30	6	4	2	1	5	32	4	1	5	4	32	31	1	5	5	36	1	5	n/a
46	-	13	26	7	-	5	-	40	1	-	10	4	32	28	16	-	5	26	16	4	-	-	28	4	14	-	46	-	n/a
-	29	2	10	17	-	21	-	8	-	-	4	19	6	3	5	-	21	3	5	21	-	-	-	20	9	-	7	32	n/a
13	2	17	-	-	-	-	1	16	-	1	7	-	9	16	-	1	-	16	-	-	1	1	14	-	2	1	15	1	n/a
26	10	-	41	-	-	7	1	31	2	1	2	7	31	10	25	-	6	10	25	6	-	-	11	6	24	3	20	8	n/a
7	17	-	-	32	-	24	4	4	-	1	11	20	-	3	1	4	24	4	1	23	4	3	4	23	2	5	9	18	n/a
-	-	-	-	-	27	-	27	-	-	27	-	-	-	-	-	27	-	-	-	-	27	27	-	-	-	27	-	-	n/a
5	31	-	7	24	-	31	-	-	-	-	4	27	-	-	1	-	30	1	1	39	-	-	1	21	2	2	8	21	n/a
-	-	1	1	4	27	-	33	-	-	30	3	-	-	-	1	32	-	-	1	-	32	31	-	1	1	33	-	-	n/a
40	8	16	31	4	-	-	-	51	-	-	12	-	39	29	22	-	-	29	22	-	-	-	28	-	23	1	45	5	n/a
1	-	-	2	-	-	-	-	-	2	-	1	-	1	-	2	-	-	-	2	-	-	-	-	-	2	-	1	1	n/a
-	-	1	1	1	27	-	30	-	-	30	-	-	-	-	1	29	-	-	1	-	29	29	-	-	1	30	-	-	n/a
10	4	7	2	11	-	4	3	12	1	-	20	-	-	11	2	3	4	12	2	3	3	2	10	3	5	4	14	2	n/a
4	19	-	7	20	-	27	-	-	-	-	-	27	-	-	1	-	26	-	1	26	-	-	-	26	1	1	6	20	n/a
32	6	9	31	-	-	-	-	39	1	-	-	-	40	18	22	-	-	18	22	-	-	-	19	-	21	1	34	5	n/a
25	3	16	10	3	-	-	-	29	-	-	11	-	18	29	-	-	-	29	-	-	-	-	25	-	4	-	28	1	n/a
16	5	-	25	1	-	1	1	22	2	1	2	1	22	-	26	-	-	-	26	-	-	-	3	-	23	3	18	5	n/a
-	-	1	-	4	27	-	32	-	-	29	3	-	-	-	-	32	-	-	-	-	32	31	-	1	-	32	-	-	n/a
5	21	-	6	24	-	30	-	-	-	-	4	26	-	-	-	-	30	1	-	29	-	-	1	28	1	1	8	21	n/a
26	3	16	10	4	-	1	-	29	-	-	12	-	18	29	-	-	1	30	-	-	-	-	26	-	4	-	29	1	n/a
16	5	-	25	1	-	1	1	22	2	1	2	1	22	-	26	-	-	-	26	-	-	-	3	-	23	3	18	5	n/a
4	21	-	6	23	-	29	-	-	-	-	3	26	-	-	-	-	29	-	-	29	-	-	-	28	1	1	7	21	n/a
-	-	1	-	4	27	-	32	-	-	29	3	-	-	-	-	32	-	-	-	-	32	31	-	1	-	32	-	-	n/a
-	-	1	-	3	27	-	31	-	-	29	2	-	-	-	-	31	-	-	-	-	31	31	-	-	-	31	-	-	n/a
28	-	14	11	4	-	1	-	28	-	-	10	-	19	25	3	-	1	26	3	-	-	-	29	-	-	-	29	-	n/a
4	20	-	6	23	-	28	1	-	-	-	3	26	-	-	-	1	28	-	-	28	1	-	-	29	-	2	6	21	n/a
14	9	2	24	2	-	2	1	23	2	1	5	1	21	4	23	-	1	4	23	1	-	-	-	-	28	3	19	6	n/a
-	-	1	3	5	27	2	33	1	-	30	4	1	1	-	3	32	1	-	3	1	32	31	-	2	3	30	-	-	n/a
46	7	15	30	9	-	8	-	45	1	-	14	6	34	28	18	-	8	29	18	7	-	-	29	6	19	-	54	-	n/a
-	22	1	8	18	-	21	-	5	1	-	2	20	5	1	5	-	21	1	5	21	-	-	-	21	6	-	-	27	n/a
n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

4.3.3 Hierarchical Cluster Segmentation (HIC)

General Illustration of Clustering Procedure

As for hierarchical clustering, Ward´s method is used. It is an agglomerative method, i.e. each observation starts out as its own cluster. Subsequently, the two closest clusters (or individuals) are combined into a new aggregate cluster. Eventually, all individuals are grouped into one large cluster. Distance between two clusters is computed as the minimum variance, i.e. the ANOVA sum of squares between the two clusters added up over all variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation (for details of the computation see SAS Institute 1994a, p. 326). The outcome is a dendrogram, i.e. a tree graph, which shows the sequence of aggregating clusters at each step (cp. Figure 24 on page 215). Additionally, a graph of subsequent differences (distances) between clusters merged at each step allows for a scree test of the possibly appropriate number of clusters to retain (cp. Table XLIV on page 216 for actual distances at each step). Sudden increases or jumps in the distance measure (analogous to error variability) suggest the appropriate number just one step before the jump in distance measure.

Parameter Choice and Rationale

Methods considered for hierarchical clustering were Average Linkage, Cetroid method, Ward´s method, Single Linkage, and Complete Linkage. Ward´s method was chosen as it seems to provide a good compromise among desirable theoretical properties, as for instance bias to join clusters with small numbers of observations, bias towards producing clusters with roughly the same number of observations, sensitivity to outliers, and the like (cp. SAS Institute 1994a, p. 326 with additional references; Hair, Anderson, Tatham, and Black 1992, pp. 274).

Resulting Clusters

In order to decide upon the appropriate number of clusters, a scree test provides some mathematical indication as to the adequate number which is inherent in this procedure, and which may be used in connection with theoretical deliberations about classes of buyers on page 195. Table XLIV on page 216 shows subsequent distance measures between clusters merged, and the graph at the bottom of the dendrogram (scree test) in Figure 24 on page 215 provides its graphical representation. The scree test shows a jump in cluster distances between the second and third cluster, suggesting three (3) clusters as the appropriate number. However, examining the pattern of joins in the dendrogram, four (4) or even five (5) clusters seem justifiable. Subsequent analyses, however, are confined to three (3) and four (4) clusters.

Figure 10 on page 203 and Figure 18 on page 207 show cluster profiles for three (3) and four (4) clusters, respectively. Those clusters exhibit clear distinctions in value attributions for certain product attributes. Using the legends on respective figures, cluster profiles may be characterized along the following lines:

c1Ward3 shows highest preference for features concerning expansibility and connectivity when considering likelihood of purchase of a notebook computer, with about equal sensitivity for price. The third and fourth attribute influencing decisions about purchase of a laptop computer may be its battery life and type of display. All other product features show part-worth magnitudes that may not be distinguished from random noise. c1Ward3 may therefore be labeled the feature-sensitives. Respondents in cluster c2Ward3 mainly seem to base their purchase decisions on the type of display and the laptop´s price, with features for connectivity and expansibility, and speed, constituting minor issues. c2Ward3 may be labeled the display-sensitives. Respondents in c3Ward3, finally, are overwhelmingly price-sensitive.

Examining cluster profiles of the 4-cluster solution in Figure 18 on page 207 and of cluster overlap with the 3-cluster solution in Table XLIII on page 211 reveals exact compatibility of display-sensitive and price-sensitive segments. However, the feature-sensitives of c1Ward3 fall into two distinct groups in the 4-cluster solution: c1Ward4 and c2Ward4. The former group may be characterized as the clearly connectivity-based feature-sensitives, whereas the latter put about equal emphasis on battery life and are also somewhat influenced in their purchase decision by keyboard size. From this discussion it may safely be concluded that Ward´s method found valid cluster solutions.

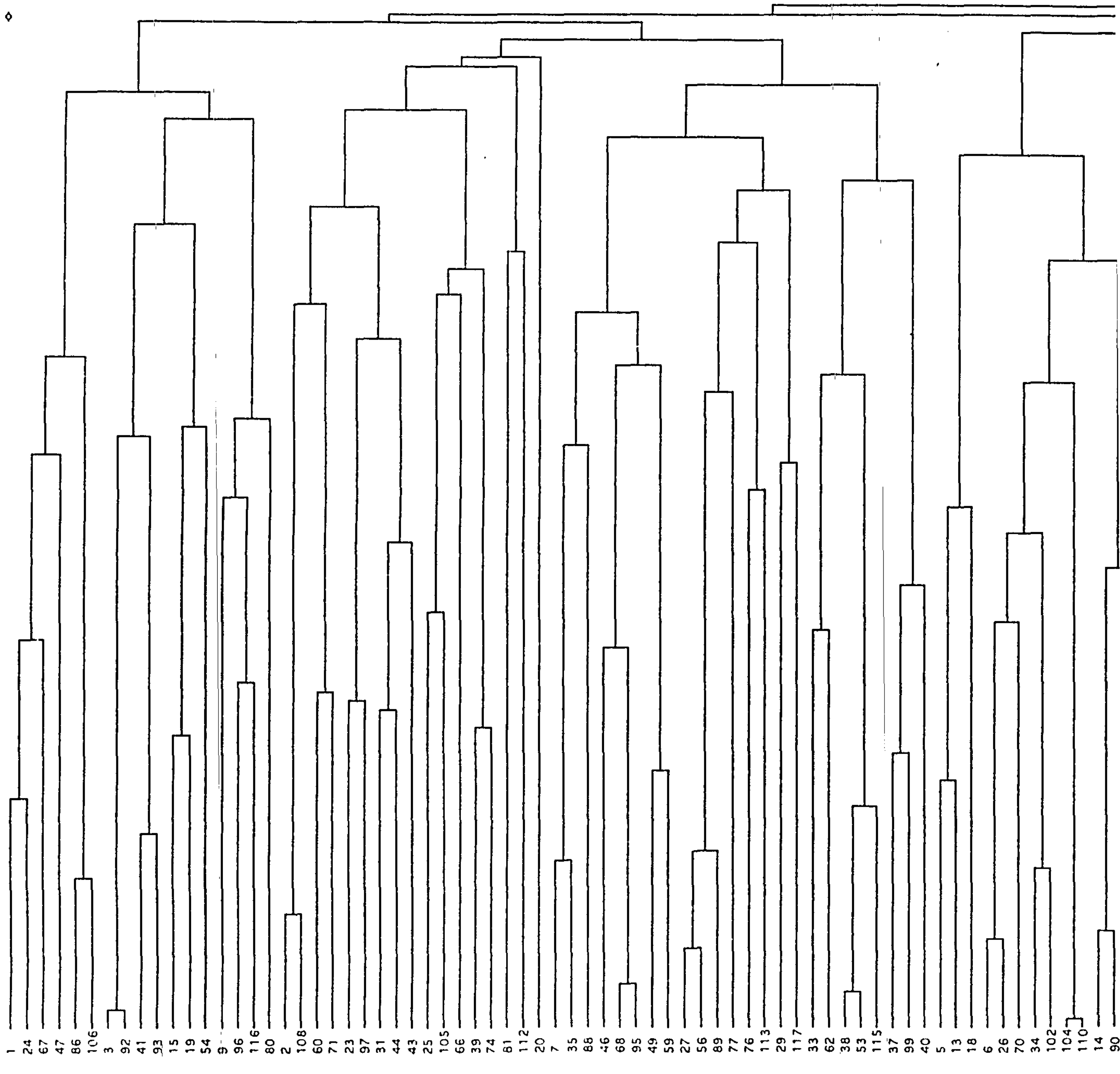


Figure 24. Dendrogram and Scree Test of Hierarchical Cluster Solution (Ward's Method).

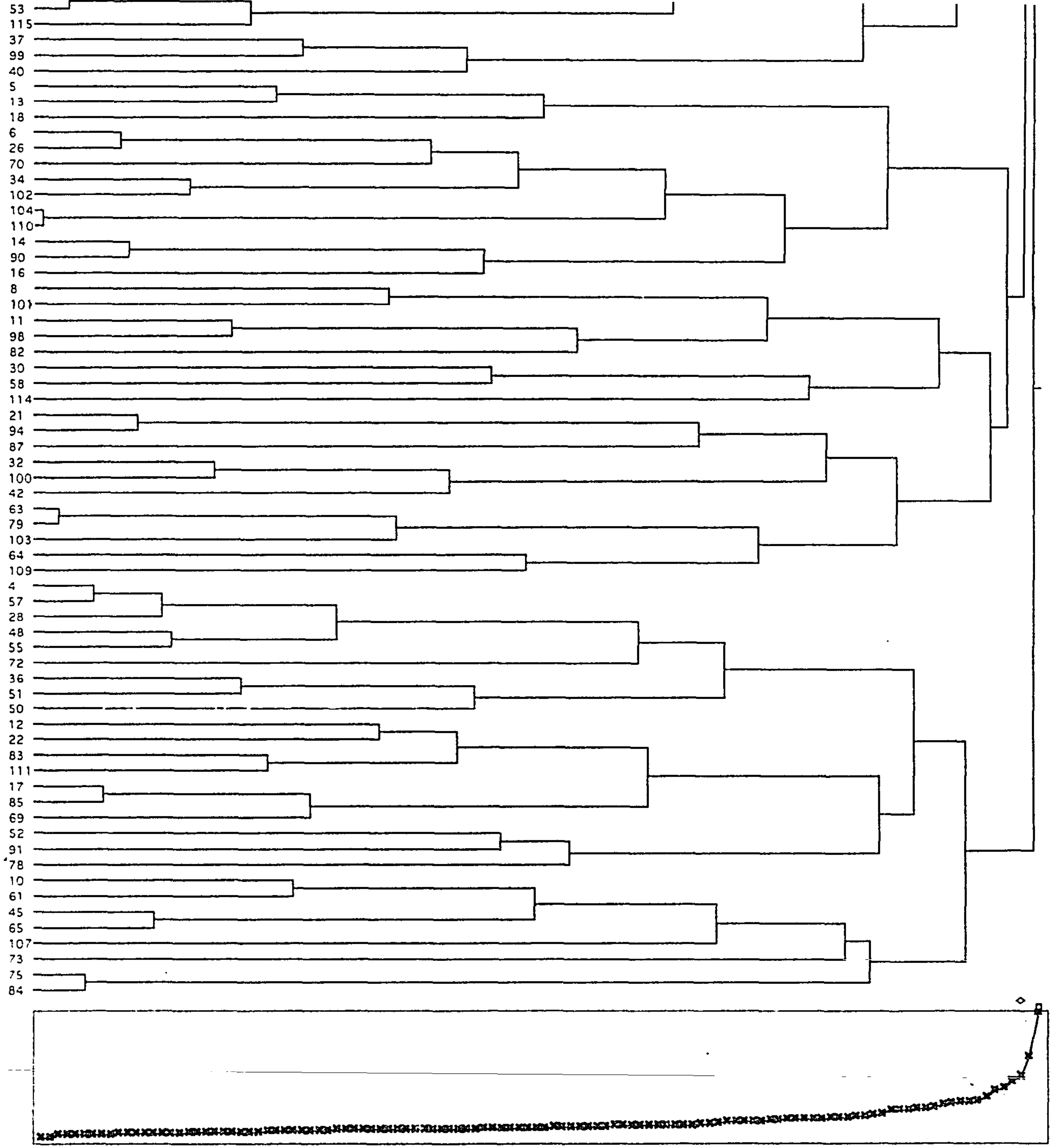

53
115
37
99
40
5
13
18
6
26
70
34
102
104
110
14
90
16
8
101
11
98
82
30
58
114
21
94
87
32
100
42
63
79
103
64
109
4
57
28
48
55
72
36
51
50
12
22
83
111
17
85
69
52
91
78
10
61
45
65
107
73
75
84

TABLE XLIV

CLUSTER DISTANCES OF HIERARCHICAL CLUSTER SOLUTION (WARD'S METHOD)

	Beginning				Intermediate				Final		
# of Clusters	Distance	Leader	Joiner	# of Clusters	Distance	Leader	Joiner	# of Clusters	Distance	Leader	Joiner
116	1.4713	104	110	77	3.0301	12	22	38	4.5869	10	107
115	1.5300	3	92	76	3.1133	8	101	37	4.6176	4	36
114	1.9440	63	79	75	3.1254	63	103	36	4.6252	7	46
113	1.9630	38	53	74	3.1284	46	68	35	4.7904	2	60
112	2.0711	68	95	73	3.1345	1	67	34	4.8352	25	66
111	2.1270	75	84	72	3.1518	33	62	33	4.8730	63	64
110	2.1293	4	57	71	3.1717	6	70	32	4.9382	8	11
109	2.1549	17	85	70	3.1768	25	105	31	5.1165	25	39
108	2.1622	27	56	69	3.1878	32	42	30	5.1839	6	14
107	2.1863	6	26	68	3.2421	12	83	29	5.2116	81	112
106	2.2683	14	90	67	3.2483	37	40	28	5.2926	27	76
105	2.2929	21	94	66	3.2564	36	50	27	5.3369	30	114
104	2.3136	2	108	65	3.2822	14	16	26	5.3602	3	15
103	2.3571	45	65	64	3.3797	30	58	25	5.5076	21	32
102	2.3655	4	28	63	3.3799	52	91	24	5.5310	2	23
101	2.4254	48	55	62	3.3999	31	43	23	5.5512	10	73
100	2.4346	86	106	61	3.4281	6	34	22	5.7233	27	29
99	2.5006	34	102	60	3.4575	64	109	21	5.8328	33	37
98	2.5010	7	35	59	3.4854	10	45	20	6.0296	10	75
97	2.5031	27	89	58	3.4894	5	18	19	6.2949	12	52
96	2.5367	32	100	57	3.5493	9	96	18	7.1031	5	6
95	2.5418	41	93	56	3.5784	76	113	17	7.1892	21	63
94	2.5766	11	98	55	3.5845	52	78	16	7.2585	7	27
93	2.6201	36	51	54	3.6528	11	82	15	7.3760	4	12
92	2.6264	38	115	53	3.6810	29	117	14	7.5863	3	9
91	2.6639	1	24	52	3.7103	1	47	13	7.8150	2	25
90	2.7284	83	111	51	3.7451	7	88	12	8.3096	8	30
89	2.7475	5	13	50	3.8026	3	41	11	8.5027	1	3
88	2.7699	49	59	49	3.8054	15	54	10	8.7125	7	33
87	2.8048	10	61	48	3.8144	9	80	9	8.7933	4	10
86	2.8340	37	99	47	3.8309	4	72	8	9.0677	2	81
85	2.8708	17	69	46	3.8443	12	17	7	9.8423	2	20
84	2.9136	15	19	45	3.9917	27	77	6	11.2453	8	21
83	2.9529	39	74	44	3.9976	6	104	5	11.8915	2	7
82	2.9880	4	48	43	4.0534	33	38	4	13.1733	5	8
81	2.9943	31	44	42	4.0564	46	49	3	14.1811	1	2
80	2.9999	23	97	41	4.0624	1	86	2	18.3086	1	5
79	3.0113	60	71	40	4.0790	21	87	1	27.7017	1	4
78	3.0291	96	116	39	4.2897	23	31				

4.3.4 Non-Hierarchical Hard Clustering (NHC)

General Illustration of Clustering Procedure

K-means clustering is applied as the non-hierarchical clustering method. Before clustering, the researcher has to pre-specify the number of clusters desired. Then, the algorithm´s first step involves selection of cluster centers or seeds with the parallel threshold procedure, i.e. a random first guess of the means of the clusters (cp. SAS Institute 1994b, p. 14; Hair, Anderson, Tatham, and Black 1992, pp. 277). Each observation, i.e. response data point, is assigned to the nearest seed. Together they form a set of temporary clusters. In the next step, the seeds are replaced by the cluster means, and once again, all observations are assigned to the nearest cluster center. This process terminates when no further changes occur any more.

Distance is computed as a simple Euclidean distance between cluster seeds or means and respective data points. Runs with distance adjusted by the sample standard deviation for a variable is not appropriate as this changes within-subject relations of part-worth utilities (cp. section 4.3.2 of this chapter). Outcomes of this clustering procedure are the cluster means, assignment of observations to respective clusters, the number of data points within a cluster, and cluster standard deviations. This k-means approach is a special case of the so-called EM algorithm, where E means Estimate (i.e. the cluster means) and M stands for maximize or minimize (i.e. assigning points to the closest clusters in this case).

Parameter Choice and Rationale

Methods considered for k-means are basically variations of finding the cluster seeds (apart from parallel threshold there are sequential threshold procedures, as well as optimizing ones; cp. Hair, Anderson, Tatham, and Black 1992, pp. 277). All k-means procedures do have a problem with smaller data tables, i.e. the results obtained can be highly sensitive to the order of observations in the data matrix, especially when clusters are not clearly separate but exhibit fuzzy, overlapping boundaries. The number of clusters to request cannot be determined by visual inspections of two-dimensional slices through the data, as it is usually done with data exhibiting less dimensions. Instead, results from hierarchical clustering (scree test) and fuzzy clustering (application of subtractive clustering procedure) were used to determine the number of clusters. In the end, three (3) or four (4) clusters were deemed appropriate from theoretical deliberations (cp. first few paragraphs of section 4.3.2 on page 195 in this chapter), and from those other cluster analyses.

Resulting Clusters

Figure 11 on page 203 and Figure 19 on page 207 show cluster profiles for three (3) and four (4) clusters, respectively. These clusters, just as those found with Ward´s method, exhibit clear distinctions in value attributions for certain product features. However, there are also some pronounced deviations. Using the legends on respective figures and comparing clusters with those obtained via hierarchical clustering, group profiles may be characterized as follows:

c3Kmean3 shows highest preference for features concerning expansibility and connectivity when considering likelihood of purchase of a laptop computer. In contrast to c1Ward3, this group shows the type of diplay as the second most influential attribute, with slightly less sensitivity for price being third. The other six attributes show part-worth magnitudes on the noise level. It is difficult to label this group the feature-sensitives, as all three (3) preferential attributes exhibit about equal magnitudes. Respondents in cluster c1Kmean3 do not exhibit clear attribute preferences, probably with the exception of price being most influential, and type of display and screen size remaining negligible. This is also in contrast to c2Ward3 which could clearly be labeled the display-sensitives. But for respondents in c2Kmean3, overwhelming price-sensitivity is very similar to c3Ward3.

Examining cluster profiles of the 4-cluster solution in Figure 19 on page 207 and of cluster overlap with the 3-cluster solution in Table XLII on page 210 reveals marked deviations between Ward´s and Kmean´s solution. c1Kmean4 roughly corresponds to c3Ward4, and may be labeled the display-sensitive segment which puts heavy emphasis on the type of display for determination of product preference. Also, the clearly price-sensitive segments are c2Kmean4 and c4Ward4. c3Kmean4 and c1Ward4 may be characterized as the feature-sensitives who lay emphasis on features for expansibility and connectivity. They differ markedly in the magnitude of part-worth utilities for features, and in sensitivity for battery life. However, both groups´ classifications seem possible. c4Kmean4 is special in that this group exhibits very high sensitivity for the keyboard size as the determinant for product preference. When examining the number of respondents belonging to that cluster with Table XLIII on page 211, however, reveals only two members. Therefore, it is doubtful if this group represents a valid cluster, or if those two respondents´ part-worths represent extremes with the resulting cluster constituting an artifact of NHC. Though not as convincing as in the HIC case, it may still be concluded that NHC Kmean method found valid cluster solutions, as they are substantively interpretable and distinct from each other.

4.3.5 Fuzzy Clustering (FUC)

General Illustration of Clustering Procedure

Fuzzy c-means (fcm) is used as the fuzzy clustering method of choice (for a number of different fuzzy clustering techniques cp Bezdek 1981; Kaufman and Rousseeuw 1990, Chapter 4). It is an extension of the (hard) k-means clustering methods. In this data clustering technique each data point belongs to a cluster to a certain degree. All degrees of membership of one specific response in respective clusters sum up to one (1), i.e. the response belongs 100% to the universe to be clustered.

Fuzzy c-means, as applied here, proceeds in an iteration loop that begins with an initial random assignment of cluster centers and subsequent respective membership grades for all observations in each of these initial clusters. Iterative updating of cluster centers and membership grades for each data point moves the cluster centers to a (local or global) minimum. The iteration is based on minimizing the (Euclidean) distance (i.e. objective function) from any given data point to a cluster center weighted by that data point´s membership grade. It terminates when either the maximum number of iterations has been reached, or when the minimum amount of improvement has not been reached between two iterations. Final output of the fuzzy c-means algorithm applied is a list of cluster centers, a fuzzy partition matrix *U* that consists of the membership grades of each data point in respective clusters, and the value of the objective function, i.e. the Euclidean distance measure in this case.

Parameter Choice and Rationale

Data: The data to be clustered are all 117 part-worth utility vectors obtained via a main-effects OLS regression procedure and subsequent adjustments.

Dimensionality of the cluster base feature vector is twenty-one (21) part-worth utilities from eight (8) attributes (cp. section 4.3.2 of this chapter).

Number of clusters: Three (3) and four (4) clusters are prespecified.

U: Final fuzzy partition matrix (or membership function matrix) is used to determine cluster membership for conjoint estimates.

m: (= exponent for the partition matrix U which controls the degree of fuzziness of the cluster solution as well as the rate of convergence of the algorithm) is set to five different values, from m = 1.05 (is equivalent to low fuzziness allowed) to m = 2.0 (high allowed fuzziness).

The maximum number of iterations is set to n = 100.

The minimum amount of improvement is set to $1x10^{-5}$.

For purposes of segment-based conjoint estimates the maximum membership value in a cluster is used for assignment of respondents to clusters, i.e. market segments. As is obvious, a different scheme for selection and assignment of data points to clusters could be used, as for instance only data points with at least 60% membership grade in a cluster could be considered distinct members of a segment.

Resulting Clusters

The number of clusters was tentatively determined with a new algorithm by Chiu (1994) called 'subtractive clustering', and compared with results of the scree test from HIC (Figure 24 on page 215). The algorithm is an extension of the 'mountain clustering' method proposed by Yager (1992). Cluster centers are estimated in a set of data assuming each data point is a potential center. A measure of the potential for each data point to be a center is calculated based on the density of surrounding data points. Then, the response with the highest measure of potential as a center is selected as the first center, and the potential of responses "near" this center is destroyed. Thereafter, the response with the next highest potential is selected and the potential of surrounding responses to become a center is destroyed. The process of acquiring a new cluster center and destroying potential "near" this response is repeated until the potential of all data points falls below a threshold. One problem with this algorithm is its use of a unit hyperbox as the clustering space. As has been explained in section 4.3.2 on page 198, this distorts within-subject relationships of part-worth utilities and invalidates respective results. To counter this effect, the algorithm was also applied without normalization, for which properties of the algorithm are not known. Cluster centers obtained with both, normalized and unnormalized, approaches did not yield valid cluster solutions. Differences among clusters tended to blur. Therefore, substantive deliberations and results of the scree test with HIC were also used to determine the appropriate number of fuzzy clusters.

Figures 12 to 16 on pages 204 to 206 show profiles of FUC cluster solutions for three (3) clusters and different allowed degrees of fuzziness with m = 1.05 to m = 2.0. Figure 17 on page 206 and Figures 20 to 23 on pages 208 to 209 illustrate profiles for FUC cluster solutions with four (4) clusters and the same different allowed degrees of fuzziness with m = 1.05 to m = 2.0. Both groups of solutions exhibit distinct differences in all respective cluster profiles found, except for solutions with the highest degree of fuzziness (m = 2.0). As for the 3-cluster solutions, substantive interpretation of cluster profiles is according to the following lines exemplified with the solution for m = 1.05:

c1fcm1.05c3 shows highest value attribution to additional features of expansibility and connectivity, followed by price and battery life, and with marginal influence of the type of display on part-worth utility values. This type of buyer may clearly be labeled the feature-sensitive. Respondents in group c2fcm1.05c3 show highest sensitivity in their preference for the type of display that comes with the notebook computer, with price exhibiting the second highest impact, and with weight coming third. This group of respondents may reasonably well be labeled the display-sensitives. Finally, segment c3fcm1.05c3 may clearly be labeled the price-sensitive one, with marginal influence of features of connectivity and expansibility on their preferences.

An important issue in fuzzy clustering is how the degree of fuzziness changes the solution, and what the best degree of fuzziness should be. In theory, higher allowed degrees of fuzziness should result in less pronounced distinctions among segment profiles. In the solutions found in this study, and with these empirical data, this effect is especially visible in the less distinct attributes when comparing fcm1.05 with fcm1.5, as for instance in battery life, speed, and weight. The highest contrasts seem to be achieved with fcm1.05 and fcm1.1, which is also slightly higher than the HIC solution in Figure 10 on page 203. Comparing Figure 16 and Figure 17 on page 206 (m = 2.0) with the other fuzzy solutions it turns out that both solutions are virtually identical and not congruent with the less fuzzy solutions. fcm2.0 for three (3) and four (4) clusters both found only two not very distinct clusters, both approaching the profile of the grand mean. These two segments very likely are invalid cluster solutions.

Turning our attention to FUC 4-cluster solutions two effects are remarkable: First, as already observed in the HIC solutions, the feature-sensitive cluster of the 3-cluster solution splits into two distinct groups while the display-sensitive and price-sensitive segments remain intact in the 4-cluster solution. With fcm1.05c4 as the example (cp. Figure 20 on page 208), c1fcm1.05c4 represents the price-sensitive segment, and c3fcm1.05c4 represents the display-sensitive one. Of the feature-sensitive groups, c2fcm1.05c4 is more influenced by the notebook´s performance/speed after features and price, while c4fcm1.05c4 is more influenced by battery life and marginally by keyboard size. Second, comparing 4-cluster solutions, profile contrasts now seem to be highest with fcm1.1 and fcm1.25, i.e. solutions which already allow a considerable degree of fuzziness. Obviously, and contrary to one´s intuition, there is no monotonic decrease in contrast among segment profiles in accord with an increase of fuzziness allowed. This finding allows for the conclusion that there is at least one optimal solution for the degree of fuzziness that optimizes contrasts among segments, and this solution need not be the one with absence of fuzziness. From this and prior paragraphs´ discussion it may safely be concluded that not-too-fuzzy FUC methods found valid cluster solutions.

4.3.6 Summary of Cluster Validity

All different clustering procedures did yield concordant clusters, i.e. clusters which are very similar in their substantive interpretation, with the exception of the non-price-sensitive clusters obtained with NHC: in this 3-cluster solution feature-sensitive and display-sensitive respondents are non-distinct; in this 4-cluster solution one cluster is comprised of only two (2) respondents who are very dissimilar to the rest. Nevertheless, with the other methods there is prevailing concordance of substantive cluster interpretation. Furthermore, cluster solutions obtained with specific methods are predominantly distinct, except for the most fuzzy solutions with m = 2.0. And finally, for the most part, different clustering methods arrived at similar partitions. Examination of those three (3) criteria lead to the conclusion that cluster procedures applied to this empirical data set resulted in valid cluster solutions. Usefulness of cluster solutions for marketing practice which is also often denominated as cluster validity is examined in more detail in the section answering Research Question # 8 on pp. 234.

4.3.7 Segment-Level Conjoint Procedures and Results

Research Question # 5.

Can cluster-based segmentation approaches improve accuracy in prediction of value attributions to product profiles over individual-level conjoint models ?

Research Question # 6.

Which aggregate model for customer value structure performs best with respect to prediction ?

As has already been stated, from the literature review about the nature of value, it was concluded one may reasonably well assume highly idiosyncratic sets of relevant attributes and model forms. This also suggests that individual-level conjoint models should outperform segment-based conjoint models in terms of predictive accuracy. However, more recent literature and pilot studies about aggregate conjoint models suggest that segment-level based methods should outperform individual-level part-worth utility models because of more stable parameter estimates, though there may be increased individual variance. This claim has not been confirmed in one replication of one particular model. Therefore, the hypothesis for this research question may be stated as follows:

H_0: Segment-level part-worth utility models do not influence predictive performance.

H_A: Segment-level part-worth utility models do influence predictive performance.

This hypothesis was first tested by performing one-way ANOVA on selected pairs of segment-level models and over selected performance measures in order to make test results comparable to individual-level models. Unfortunately, not all desirable tests could be conducted due to violations of test assumptions. Therefore, in addition to the ANOVA, paired t-tests and Chi-Square tests were conducted.

Estimation

After obtaining valid cluster solutions, these were used to estimate segment-level conjoint models according to Figure 7 on page 88. Table XLV on page 227 lists performance measures calculated for segment-level models. All figures are weighted averages of respective cluster solutions. The mean count of First-Hit which was calculated for the individual-level models is missing for the segment-level models, as it is not used for tests, here.

When examining Table XLV on page 227 the most obvious result is that none of the segment-level measures exceeds individual-level performance measures, neither for the conjoint, nor for the self-explicated models. On the contrary, and except for First-Hit, all performance measures are much lower in absolute values than their individual-level counterparts. Another important observation is that best and worst model performance is dependent on the measure used for the comparison. This is unfortunate, as it limits generalizability of model performance and the usefulness of associated tests of significance. Yet another dilemma is the absence of tests (or the violation of test assumptions) for most performance measures to determine relative performance of models. For this problem, a Monté Carlo study could determine levels of confidence, significance, and power of differences in performance for different performance measures and selected segmentation-based conjoint models. One such attempt is the Monté Carlo study of Vriens, Wedel, and Wilms (1992) which, regrettably, is not useful for interpretation of the present study due to missing parameter variations. Umesh and Mishra's (1990) Monté Carlo study for R^2 is not applicable for segment-level conjoint models.

TABLE XLV

PREDICTIVE PERFORMANCES OF SEGMENT-LEVEL MODELS

	Type of Model	Method	# of Clusters	Performance Measures (Averages Over Groups)							
				R^2 (calib.) [b]	Adj R^2 (calib.)	R^2 (hold.) [b]	r_{xy} (hold.) [a]	Fisher's $z(r_{xy})$ (hold.) [a]	RMSE (hold.)	RMSE (calib.)	First-Hit (hold.)
1.	HIC:	Ward	3	0.3102	0.3007	0.3126	0.5538	0.6307	26.02	23.20	0.6453
2.	NHC:	Kmean	3	0.3007	0.2905	0.2955	0.5408	0.6088	26.40	23.57	0.6538
3.	FUC:	m = 1.05	3	0.3139	0.3042	0.3152	0.5573	0.6342	25.99	23.27	0.6560
4.	FUC:	m = 1.1	3	0.3139	0.3042	0.3152	0.5573	0.6342	25.99	23.27	0.6560
5.	FUC:	m = 1.25	3	0.3108	0.3010	0.3139	0.5573	0.6327	26.06	23.36	0.6538
6.	FUC:	m = 1.5	3	0.3096	0.2996	0.3162	0.5600	0.6359	25.95	23.38	0.6410
7.	FUC:	m = 2.0	3	0.3017	0.2917	0.3052	0.5490	0.6211	26.19	23.59	0.6453
8.	HIC:	Ward	4	0.3264	0.3137	0.3165	0.5573	0.6358	25.98	22.94	0.6517
9.	NHC:	Kmean	4	0.3286	0.3173	0.3291	0.5688	0.6549	25.64	22.92	0.6645
10.	FUC:	m = 1.05	4	0.3314	0.3185	0.3133	0.5537	0.6311	26.01	22.89	0.6709
11.	FUC:	m = 1.1	4	0.3319	0.3189	0.3216	0.5639	0.6429	25.91	22.99	0.6752
12.	FUC:	m = 1.25	4	0.3325	0.3195	0.3238	0.5654	0.6458	25.81	22.93	0.6752
13.	FUC:	m = 1.5	4	0.3305	0.3175	0.3202	0.5611	0.6409	25.89	22.92	0.6816

2.	NHC: Kmean	3	0.3007	0.2905	0.2955	0.5408	0.6088	26.40	23.57	0.6538
3.	FUC: m = 1.05	3	0.3139	0.3042	0.3152	0.5573	0.6342	25.99	23.27	0.6560
4.	FUC: m = 1.1	3	0.3139	0.3042	0.3152	0.5573	0.6342	25.99	23.27	0.6560
5.	FUC: m = 1.25	3	0.3108	0.3010	0.3139	0.5573	0.6327	26.06	23.36	0.6538
6.	FUC: m = 1.5	3	0.3096	0.2996	0.3162	0.5600	0.6359	25.95	23.38	0.6410
7.	FUC: m = 2.0	3	0.3017	0.2917	0.3052	0.5490	0.6211	26.19	23.59	0.6453
8.	HIC: Ward	4	0.3264	0.3137	0.3165	0.5573	0.6358	25.98	22.94	0.6517
9.	NHC: Kmean	4	0.3286	0.3173	0.3291	0.5688	0.6549	25.64	22.92	0.6645
10.	FUC: m = 1.05	4	0.3314	0.3185	0.3133	0.5537	0.6311	26.01	22.89	0.6709
11.	FUC: m = 1.1	4	0.3319	0.3189	0.3216	0.5639	0.6429	25.91	22.99	0.6752
12.	FUC: m = 1.25	4	0.3325	0.3195	0.3238	0.5654	0.6458	25.81	22.93	0.6752
13.	FUC: m = 1.5	4	0.3305	0.3175	0.3202	0.5611	0.6409	25.89	22.92	0.6816
14.	FUC: m = 2.0	4	0.3022	0.2923	0.3112	0.5537	0.6290	26.06	23.59	0.6453

(calib.) = Calibration set
(hold.) = Holdout set

[a] = Seemingly non-monotone transformations between r_{xy} and z (r_{xy}) when comparing different cells in the table result from averaging individual results which is appropriate for Fisher's z, but not for r_{xy} .

[b] = Some marginally better R^2 for the holdout rather than for the calibration set of profiles results from different methods for calculation: as for the holdout set, r_{xy} was squared first, then the weighted average was computed for the respective cluster method, while R^2 for the calibration set of profiles was computed directly for the clusters, then the weighted average for the method was obtained.

All figures are weighted averages of respective segment-level numbers.

R^2, the variance accounted for in the calibration set of profiles is only a bit above 30% for the worst model ($R^2 = 0.3007$ for Kmean3), and at 33.25% for the best model (fcm1.25c4). This is markedly worse than individual-level conjoint models for treatment groups in Phase I of this study (cp. for instance Tables XIV and XXV on pages 141 and 162). It is also markedly worse than the average R^2 (model fit) for TC main effects models over all 117 respondents (Ø $R^2 = 0.8584$). But it is still in the range of many conjoint studies which also did not yield higher "goodness-of-fit". The difference between best and worst model is over three (3) percentage points, i.e. about 10% from the worst "goodness-of-fit". Another interesting observation is the fact that most 4-cluster solutions exhibit higher performance in prediction than 3-cluster solutions, though the difference is below the three percentage mark. A possible reason may be that 4-cluster solutions better reflect differences in value attributions to product profiles than 3-cluster solutions, though both solutions are valid in terms of substantial interpretation.

Some R^2s for the holdout set of profiles are marginally better than those for the calibration sets. This effect most likely does not reflect overfitting, but may be explained with the difference in the way both R^2s were computed: For the holdout set of profiles, r_{xy} was squared first, then the weighted average was computed for the respective cluster method, while R^2 for the calibration set of profiles was computed directly for the clusters, then the weighted average for the method was obtained. As the scale of r_{xy} is not an interval scale (see page 120 and footnote 20 on page 143), these different approaches lead to different results even with the same cluster solution and data set. The worst R^2 for the holdout data is $R^2 = 0.2905$ for Kmean3, and the best $R^2 = 0.3291$ for Kmean4. Neglecting Kmean4 due to the doubtful validity of its cluster solution, the best holdout R^2 is $R^2 = 0.3238$ for fcm1.25c4. As for Adj R^2 for the calibration set, Kmean3 performed worse with Adj $R^2 = 0.2905$ and fcm1.25c4 performed best with Adj $R^2 = 0.3195$. This is also much worse than average adjusted R^2 for all individual-level TC main effects conjoint models (Ø Adj $R^2 = 0.6654$).

With the holdout set of profiles, r_{xy}, Fisher's z(r_{xy}), and RMSE show the same pattern of performance: Kmean3 is worst. Kmean4 is best; when neglecting Kmean4 due to doubtful validity of its cluster solution, fcm1.25c4 is the best segment-level conjoint model.

RMSE for the calibration set of profiles shows both most fuzzy models being worst, i.e. fcm2.0c3 and fcm2.0c4 with RMSE = 23.59 which, however, is only slightly worse than Kmean3 (RMSE = 23.57). Best calibration RMSE is exhibited with 22.89 by fcm1.05c4. Predictive performance with the holdout set of profiles in terms of First-Hit also shows quite a different pattern: fcm1.5c3 performs worst with 64.1% correctly predicted first hits, and fcm1.5c4 performs best with 68.2% correctly predicted first hits. Only First-Hit performance measure approaches magnitudes reached with individual-level main effects conjoint models.

In summary, observing different performance measures and the differences between selected segmentation methods, it seems that segment-level conjoint models on the basis of Kmean3 perform worst, and models using fcm1.25c4 perform best. It is remarkable that fuzzy models have the ability to perform best, but it seems to depend on the degree of fuzziness allowed. Finally, as patterns of performance across models and performance measures are not unambiguous, it is important to know if at least differences between best and worst model show significance. This test is done next.

Possible Tests

Comparing segment-level conjoint models with each other entails problems absent with individual-level models. First, ANOVAs with Fisher's z(r_{xy}) and RMSE measures cannot be performed any more due to insufficient numbers of data points. Each segment provides for one number, i.e. performance measure, leading to only three and four measures for the 3-cluster and 4-cluster solutions. Second, more direct tests for comparisons between Pearson product moment correlations r_{xy} as suggested on pp. 120 are not possible because segment-based model r_{xy}s do not exhibit only binomial distributions, and variances as well as the number of respondents in each segment (i.e. the basis of each proportion) are very different for each segment. Third, within the same lines of arguments, a direct test for First-Hits between segment-based models is not feasible because distributional assumptions of the z-tests on pp. 121 for First-Hit are not met.

Nevertheless, there is the possibility to test differences in performance among segment-based conjoint models on the basis of First-Hit counts per respondent. From an inspection of the absolute values of performance measures, and even without formal tests, one may easily conclude for Research Question # 5 that cluster-based segmentation approaches cannot improve accuracy in prediction of preferences and choice behavior vs. individual-level conjoint models. Backing this claim with formal tests of performances between individual-level and segment-level models revealed impossible as some prerequisites for valid tests are absent, especially the assumption of normal distribution of the data (e.g. for counts of First-Hit for the holdout set of profiles, and fcm1.5c3, Shapiro-Wilk $W = 0.8737$; $p < 0.0000$), and a sufficient number of data points to perform valid statistics with segment-level models.

In order to answer Research Question # 6, and in order to apply a test that, theoretically, allows for comparison of individual-based and segment-level conjoint models, an ANOVA test and a paired t-test is performed for First-Hit performance measure with the holdout set of profiles, and between best and worst performing segmentation methods over 3-cluster and 4-cluster solutions, as well as for 4cluster solutions, alone. However, as the distributional assumption of normality is not met for this measure, a Chi-Square test as a test of independence is provided in addition to the former two tests, though the latter statistic provides no information about the strength or direction of the association between First-Hit measures of two segment-based conjoint models.

Results

Table XLVI on page 232 illustrates results for all three tests performed on the (over all segmentation methods) worst and best performing segment-level conjoint models, i.e. fcm1.5c3 with 64.1% correctly predicted first hits, and fcm1.5c4 with 68.2% correctly predicted first hits.

The ANOVA performed on First-Hit could not determine significance of differences with $F = 1.7373$ and $p < 0.1888$, however the paired t-test could with a t-Ratio of 2.2957 and a p-value of less than 0.0117. The Chi-Square test also yielded clear significance of differences between those two model forms. These results must be interpreted with caution as they were obtained with a data set that does not satisfy the assumption of normal distribution of responses.

TABLE XLVI

TESTS BETWEEN SEGMENT-LEVEL MODEL FORMS

Tests for First-Hit Type of Test	Model 1 vs. Model 2	Assumption of normal distribution valid ?	Test Ratio	p-value
ANOVA $F_{1,232}$	fcm1.5c4 vs. fcm1.5c3	No	F = 1.7373	0.1888
	fcm1.5c4 vs. fcm2.0c4	No	F = 1.2473	0.2652
Paired t-test (one-tailed)	fcm1.5c4 vs. fcm1.5c3	No No	t = 2.2957	0.0117
	fcm1.5c4 vs. fcm2.0c4	No No	t = 1.9102	0.0293
Chi-Square$_{12,101}$	fcm1.5c4 vs. fcm1.5c3	n/a	LL Ratio = 109.85	0.0000
	fcm1.5c4 vs. fcm2.0c4	n/a	LL Ratio = 157.35	0.0000

Tests are also ordered in terms of sensitivity for detection of differences. The ANOVA is the least sensitive test as group variances are pooled. The t-test is a stronger test of differences between groups as the variance for both groups is not pooled, but calculated separately. The Chi-square test is the most sensitive concerning the high number of responses.[29] Best and worst models´ performance is only five (5) percentage points apart, or about 6.3 % measured from the worst model. Considering the number of respondents (117), this difference may be indicative of systematic

[29] Significance of Chi-Square with large numbers of respondents is problematic as the magnitude of Chi-Square is dependent on the number of respondents. Measures that adjust Chi-Square for the number of respondents, however, as for instance phi, the coefficient of contingency C, or Cramér's V, do have problems of their own (Norusis/SPSS 1993, pp. 208).

deviations between the two models. The F-test could not detect this difference. Separating individual variance with a paired t-test, the difference between models becomes apparent. Though usually violations of the assumption of normal distribution of responses reduces power of the test, it is not clear if this phenomenon is responsible for the F-test´s nonsignificance. The Chi-Square test, in contrast, seems to be too sensitive with respect to the absolute differences between models.

Considering best and worst segment-level conjoint model for the 4-cluster solution only, differences between models are a little less pronounced: best model remains fcm1.5c4 with 68.2 % correctly predicted first choice hits, while worst model fcm2.0c4 with 64.5 % correctly predicted first hits is only 3.7 percentage points worse.

The ANOVA performed for these models´ difference was not significant with $F = 1.2473$ and $p < 0.2652$, but the paired t-test was significant, again, with a t-Ratio of 1.9102 and a p-value of less than 0.0293. The Chi-Square test yielded clear significance, too, but this result must be viewed with caution as it was obtained with a sample that is relatively large for Chi-Square measure.

In sum, one may conclude that there is a significant difference between best and worst segment-level conjoint models. Therefore, H_0 must be rejected and H_A be believed: Segment-level part-worth utility models <u>do</u> influence predictive performance. Best and worst model are the 4-cluster and 3-cluster solution of fuzzy clustering with $m = 1.5$, respectively. The 4-cluster solution seems to be able to more accurately reflect different value attributions to product profiles, leading to higher performance measures.

4.3.8 Plausibility and Practicality

Research Question # 7.

Are the purposes of prediction and segmentation, as well as potential other purposes, better served with the suggested methods, and what practical limitations are there for the different methods to support specific purposes ?

Research Question # 8.

Are benefit segments obtained with different clustering procedures meaningful for target marketing, or may they only increase predictive accuracy ?

These two questions do not lend themselves to hypothesis testing. They concern the benefit cluster solutions obtained, and possible conflicts from high predictive accuracy but poor ways to meaningfully address segments with various business policies. Ultimately, these two research questions concern usefulness of applied methods for purposes of increased accuracy in prediction, and improved segmentation.

This possible conflict did not materialize: Segment-level main-effects-only conjoint models were markedly inferior to individual-level main-effects-only conjoint models over all performance measures considered for the comparison (Table XLV on page 227). Increasing reliability of part-worth utility estimates, i.e. value structure, by trading high variance in respondents´ part-worth estimates did not simultaneously increase accuracy in prediction. Therefore, segment-level conjoint models may not be considered useful for to increase accuracy of prediction. This contradicts assumed reversal of the best model with a change from individual-level to market conjoint models as suggested with (Hagerty´s) theory and a (very limited) Monté Carlo study by Hagerty (1986, pp. 301 and 309).

Turning attention to the eighth research question first, the second part of Research Question # 8 has already been answered: Segment-level conjoint models do not increase accuracy in prediction. As for the first part of this question, validity of cluster solutions has also been established. However, meaningfulness of benefit segments obtained with different clustering procedures for target marketing refers to usefulness of solutions which is also often termed cluster validity. Meaningfulness of segments is judged upon criteria of substantiality, actionability, and accessibility.

Segments obtained with clustering procedures show distinct discriminatory level-utilities, i.e. one element of substantiality, with the exception of non-hierarchical k-means method and the most fuzzy c-means methods with m = 2.0 (Figures 11, 16, 17, and 19 on pages 203, 206, and 207). Stability of segment profiles over time, the second component of substantiality, is difficult to determine as segments change with repeated application of clustering procedures. Conceptually, it is not clear if (repeated) within-subjects or inter-subjects segment profiles are indicative of stability over time. However, one indication of stability of value structure over time has been provided with comparisons of repeated measurements: The two unconfounded comparisons of individual-level part-worth utilities after first and second measurement (for groups G3 and G5; cp. Tables XXXIV on page 176 and XXIII on page 157) yielded stability of value structure over time.

A similar consideration of stability of segment profiles related to the segmentation procedure concerns similarity of value structure for (average) individual-level part-worth utilities and part-worths obtained with segment-level conjoint estimates. Figures 25 to 27 on pages 236 and 237 show average individual-level part-worth utilities before segment-level conjoint estimates for best and worst segment-level models as tested in Table XLVI (i.e. models fcm1.5c3, fcm2.0c4, and fcm1.5c4).

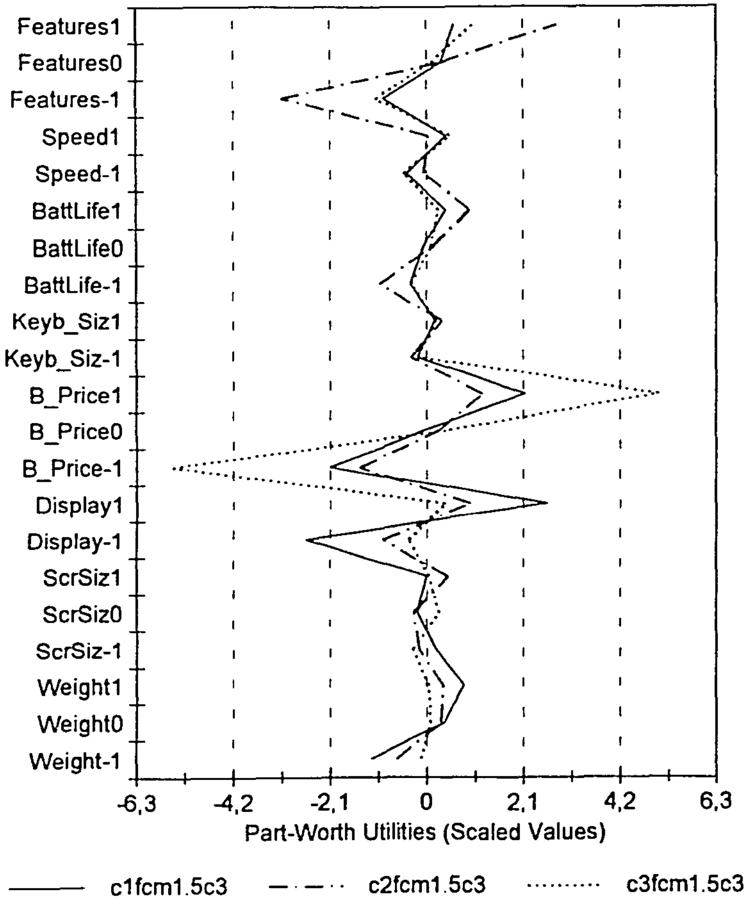


Figure 25. A Priori FUC Cluster-Profiles for 3 Clusters with m = 1.5.

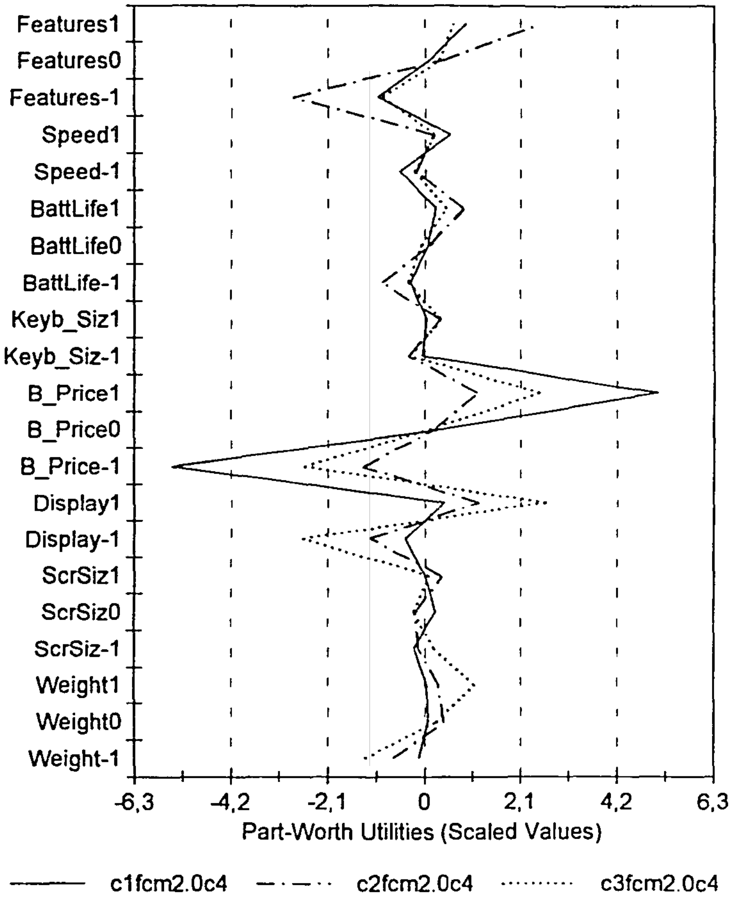


Figure 26. A Priori FUC Cluster-Profiles for 4 Clusters with m = 2.0.

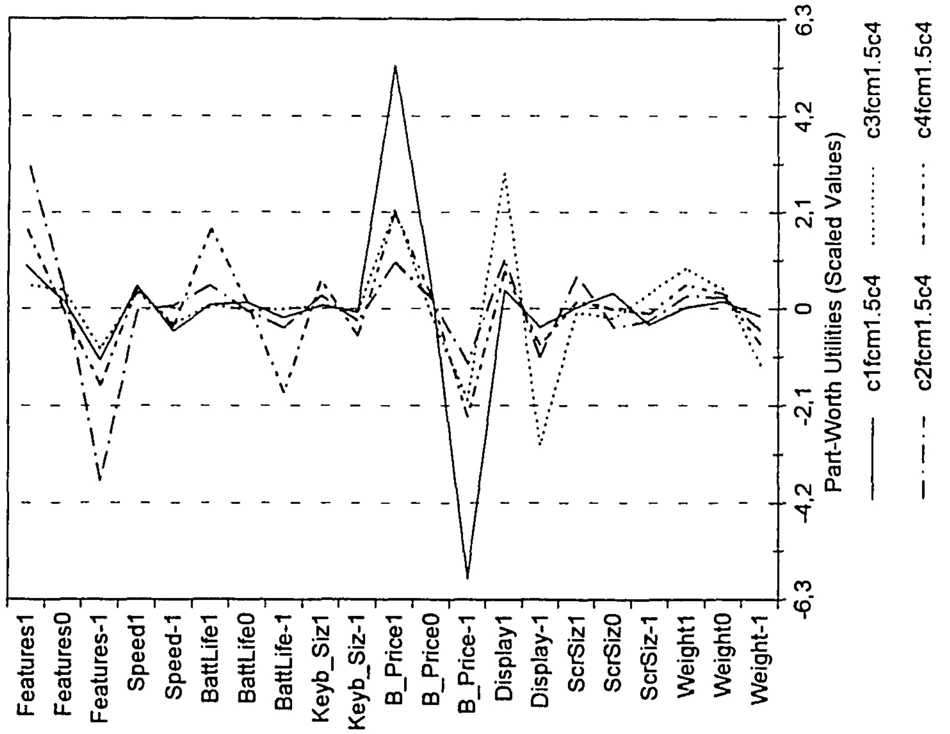


Figure 27. A Priori FUC Cluster-Profiles for 4 Clusters with m = 1.5.

Comparing a priori FUC cluster profiles for the worst-performing conjoint main effects model with three (3) clusters and m = 1.5 in Figure 25 on page 236 with the according segment-level cluster profiles in Figure 15 on page 205 there is, by and large, congruence between respective cluster profiles. Only slight deviations between individual-level and segment-level profiles in the type of display, in battery life, and in features are noticeable. Differences are grave with the most fuzzy models and m = 2.0: The most fuzzy segment-level models cannot differentiate among three cluster centers any more, but recognize only two largely similar profiles. As for FUC cluster profiles for four (4) clusters with m = 1.5, the a priori cluster profiles exhibit larger deviations among each other than the segment-level models.

From presence of discriminatory preferences, stability over time, and stability over individual-level and segment-level part-worth estimates, resulting segments reveal as substantiated and meaningful for the market researcher.

In addition, meaningfulness, and certainly usefulness of groupings of potential market participants resulting from segment-level conjoint estimates may be judged with the ability of a firm to act upon knowledge of benefit attributions with combinations of product attributes and other variables of the marketing mix. As relevancy of attributes for value judgments of respondents and relevancy of attributes for managerial actions provide the basis for conduct of a conjoint study, consideration of this issue at this point constitutes an a posteriori check of an a priori balanced study design. Evaluating segment profiles, it seems possible for a firm to provide market offerings specifically geared to market segments obtained with this study.

Finally, a criterion that may be considered part of meaningfulness, but certainly a component of usefulness, is accessibility or reachability of individuals within a specific segment. Considering the segments obtained and their clear distinctions among profiles which also facilitated labeling, it is very likely that those segments are accessible with specific product offers, and an according communication policy.

In order to maximize efficiency of access to specific segments it would be helpful to establish covariation of benefit attributions to product profiles with demographic and / or psychographic characteristics of respondents. However, due to the purpose of this study and the limited ancillary measures gathered about respondents, such an exploration could not be performed within this research study. For instance, covariation of segments with familiarity would allow to adapt communication to market participants´ product knowledge, potentially increasing efficacy of communication, and efficiency of access to the market via selected media. Selected analyses of covariation of familiarity, work experience, and other ancillary variables about respondents with segments did not yield insightful correlations.

Nevertheless, from the clear differences in part-worth utilities exhibited by the price-sensitive, the feature-sensitive, or the display-sensitive segments, one may conclude that meaningful and useful leverages for access are possible with segment-level conjoint models, but segment-level estimates do not reveal to be better than individual-level ones.

In conclusion, considering substantiality, actionability, and accessibility, segment-level conjoint models are not better, but just as good as individual-level conjoint models in determining segment-level part-worth utility profiles, i.e. value structure. Segment-level conjoint models, however, are markedly worse than individual-level models in predicting preferences and choice behavior. Furthermore, limitations of study design are nearly as grave as with individual-level models, with the exception that segment-level conjoint estimates possibly necessitate less profiles to be evaluated by any one individual. Finally, another important limitation of segment-level part-worth estimates is a lack of valid tests for preference measures r_{xy}, Fisher´s z, and RMSE (at least in this study, and with these data´s distributions) which, in contrast, are possible with individual-level estimates.

CHAPTER V

FINDINGS AND CONCLUSIONS

This chapter summarizes and expands on findings in the results section, and it discusses and interprets Phases I and II. First, major findings are detailed. Then, the contribution of systems science to this study is elaborated on. Next, contribution of this study to marketing theory and practice are illustrated. Finally, remaining limitations, and directions for future research are commented on.

5.1 Major Findings

Maybe the most general finding concerns the question if it is even worthwhile to study conjoint methodology. Without any doubt one may be assured, conjoint analysis is a method for measurement of customer value that is well worthwhile to be studied. This statement may already be obvious from the prior chapter, but may become even more so in subsequent paragraphs.

Convergent Validity and Reliability of Individual-Level Models

In order to determine convergent validity of conjoint methodology as outlined in section 2.6 (pp.68), traditional individual-level conjoint models (TC) were compared to self-explicated (SE) models for customer value. A summary for selected results is provided in Table XLVII on page 242. Consistently, the best conjoint models (TC) yielded substantial improvements in the accuracy of prediction versus the self-explicated (SE) models. For Pearson´s product moment correlation coefficient r_{xy} the improvements of TC models vs. SE models are in the magnitude of about nine to ten percentage points (9% - 10%) for three quarters of respondents in the second measurement, and at about four percentage points (4%) for all respondents in the first measurement. Regarding Fisher´s z (r_{xy}) which is more appropriate for comparisons due to its interval scale, average performance advantages for conjoint models are in the range of fourteen percent (14%). For First-Hit, conjoint models are, on average, about four (4) percentage points better than SE models. Conjoint models, on average, can account for about seven to eight percentage points (7% - 8%) more of the variance in the responses of potential customers than SE models. The important observation to establish convergent validity, however, is the fact that performance measures of SE models improved and deteriorated in accord with the conjoint models for different methodological variations, but with one exception: repeated measurements for selected performance measures of group G5 as detailed in the results section.

In terms of value structure, and for purposes of segmentation, individual-level conjoint models do have better discriminating power between attributes which becomes apparent with larger differences in attribute importances, while direct questioning for self-explicated (SE) models yielded more average importances which are hardly to distinguish from a random model (cp. Table XIII on page 137). Furthermore, part-worth utilities of conjoint models are signed which allows for an evaluation of positive and negative contribution of specific attribute levels to overall utility which is absent in self-explicated models, making it harder to interpret results. This finding is consistent with the statement that derived value attributions to product descriptions are more accurate than directly elicited ones.

TABLE XLVII

SUMMARY OF DIFFERENCES BETWEEN TC MAIN EFFECTS MODELS AND SE MODELS FOR SELECTED PERFORMANCE MEASURES AND GROUP COMPARISONS

Measurement Groups	r_{xy}	Fisher's $z(r_{xy})$	First-Hit	R^2	Details in
G5	0.0364	0.1425	0.	0.0696	Table XIV; p. 141
G6	0.0912	0.1693	7.50	0.1107	Table XIV; p. 141
G2 (G5)	0.0219	0.0978	2.50	0.0293	Table XVI; p. 146
G2 (G6)	0.0314	0.0658	7.50	0.0386	Table XIX; p. 150
G3	0.0915	0.2580	5.55	0.1388	Table XXV; p. 162
G4	0.1050	0.1878	6.67	0.1138	Table XXV; p. 162
G1 (G3)	0.0239	0.0710	-1.85	0.0341	Table XXVII; p. 165
G1 (G4)	0.0640	0.1285	2.50	0.0664	Table XXX; p. 170
Average	0.0582	0.1401	3.80	0.0752	

All numbers refer to the holdout set of profiles.
Group differences are in chronological order.

In addition, individual-level conjoint models seem to be more reliable over time than SE models. For all repeated measurements over the majority of performance measures, and for the conjoint models, accuracy is improving with the second measurement while for SE models and for two groups (G5 and G6), the second measurement seems to yield no better, and even worse results than the first. Familiarity with the task should provide assurance against deterioration in the measurement which is visible with conjoint models, and should reduce perception of difficulty of the task. Instead, an ancillary variable collected for both repeated measurements, perceived difficulty of the task, deteriorated slightly from an average of 2.57 for the first mesurement to an average of 2.62 for the second measurement out of a range from one (1) to seven (7) categories. This difference in perceived difficulty of the task, however, is not significant in a paired t-test with both measurements´ nonnormal data sets ($p < 0.3138$), but is significant with a Chi-Square test (LL $\text{Ratio}_{36,75} = 93.23$ at $p < 0.0000$). It seems that conjoint analysis gains from task familiarity while self-explicated models do not seem to be influenced by task familiarity or perceived difficulty of the task. SE models do not seem to gain in accuracy of prediction with repeated measurements, but conjoint models do, especially when performance levels are low and the respondent task is difficult.

Concerning reliability over attribute set and over stimulus set, conjoint models are also superior to SE models for both of these methodological variations. The inclusion of user-referent attribute sets is able to improve accuracy in prediction. Group comparisons showed a consistent improvement in predictive accuracy with the inclusion of a user-referent attribute into the attribute set for conjoint models. Nevertheless, this tendency was generally not statistically significant. For Fisher´s z and RMSE, H_0 could be rejected at the $\alpha < 0.1$ level, but not at the $\alpha < 0.05$ level. SE models did also show this consistency of improvement in prediction, but did not reach the magnitudes in performance of conjoint models. Quite contrary, it is hard to detect any effect from different fractional factorials at all. Conjoint analysis is very reliable over stimulus sets. This also alleviates concerns voiced in Reibstein, Bateson, and Boulding (1988, pp. 280) about possible problems with fractional factorial designs. Problems in their study may be explained with attribute interactions which they could not model and test for, but which were included in this study. Additionally, the current study found an exactly reversed effect from their study: here, reliability over stimulus set is higher than reliability over attribute set which may be explained with perturbation of only the least important attributes in their study. Finally, there is evidence that influences from different methodological variations can cancel out when those are combined which point in opposite directions, thus increasing overall reliability of the method.

Apart from statistical considerations but in contrast to the positive empirical properties of increased accuracy, user-referent attributes pose the problem of possible ambiguities in understanding among respondents, making it more difficult, in practice, to attach the beneficial attribute to one´s offer. For example, firm reputation may mean different things to different potential customers. On the other hand, user-referent attributes as firm reputation allow for a measurement of decision (i.e. evaluative) criteria that are more comprehensive than simply the (physical) product offer. And the important finding, here, is conjoint analysis´ ability to measure such influences on potential customers´ preferences and choices.

Segment-Level Performance

Segment-level conjoint estimates performed much worse in terms of prediction than individual-level conjoint models. They also were not perceivably better in exhibiting value structure than individual-level models. Therefore, segment-level conjoint estimates cannot be recommended if violation of statistical prerequisites for conjoint analysis can be avoided with appropriate planning of the conjoint study. However, clustering after estimation of individual-level conjoint models can be recommended as an effective means for exhibition of value structure of possible market segments.

One conjecture why individual-level models came out better in this study than segment-level models is that the individual-level models already leave sufficient degrees of freedom for error (11 and 9 in this study). Therefore, bias in the parameter estimates may not be an issue, here, as parameter estimates are already very stable. Considering, for example, one study with similar numbers in the degrees of freedom, and the most extensive study to date that takes Bateson, Reibstein, and Boulding's (1987) framework for conjoint reliability into account, i.e. distinguishes between reliability and validity, as well as among different forms of reliability, their study leaves between nine (9), and eighteen (18) degrees of freedom for error for varying products and numbers of attribute levels included in the study (Reibstein, Bateson, and Boulding 1988, p. 276), and also establishes high reliability for individual-level conjoint models. However, each of their group comparisons is based on only 20 respondents per cell, and their measure of reliability, the alpha level resulting from a Chow test (specific F-test) of the possibility to pool test applications, has been shown to increase when the number of product profiles decreases which is exactly the opposite of what one should expect in a reliability measure (Green and Srinivasan 1990, p. 12). This study avoids this measure, and also avoids reliance on only one performance measure. Instead, this study's findings are based on a total of nine (9) performance measures where (admittedly) some are related, and on up to three different tests. These measures were computed over a holdout set of sixteen (16) profiles. This approach should provide greater confidence in the study's findings than in those of earlier studies.

From this study's results it may be concluded that individual-level models are best when some basic statistical requirements are met, as for instance leaving sufficient degrees of freedom for error, and basing performance measures on a sufficient number of holdout profiles. Umesh and Mishra, based on their Monté Carlo study, also regard "the residual degrees of freedom of the conjoint analysis design" as "the most important factor that influences the goodness of fit" (1990, p. 43). In addition, when statistical requirements are met, the gain of reduced bias from segment-level conjoint procedures seem to be outweighed by the increased variance of individual respondents, leaving the performance advantage with individual-level models.

This study's finding of superior individual-level conjoint models is also in line with the only two other limited empirical studies that compared individual-level and segment-level conjoint models (Green and Helsen 1989; Green, Krieger, and Schaffer 1993a). In both these replications, however, degrees of freedom were not always higher than in the original studies, neither were performance measures always based on a higher number of holdout profiles: Green and Helsen (1989) used eighteen (18) calibration profiles, and sixteen (16) validation profiles (holdout set). This left only five (5) degrees of freedom for the calibration set. However, Hagerty's study (1985) claiming superiority of segment-level models, used sixteen (16) calibration profiles and only two (2) holdouts. This also left five (5) degrees of freedom for calibration, but performance is based on only two (2) responses which may (at least partly) be responsible for bad results of individual-level models. Two other possible reasons for failure to replicate Hagerty's findings may be his usage of ranking (i.e. ordinal) data as rating (i.e. interval) data, and standardization of responses of individuals before exposition to the Q-type factor analysis. Kamakura's study (1988) used twenty-seven (27) calibration and eight (8) holdout profiles, leaving sixteen (16) degrees of freedom for calibration, but performance measures are also based on fewer holdout profiles, and statistical significance is assumed at the $\alpha = 0.1$ level. Green, Krieger, and Schaffer's study (1993a) used, respectively, three data sets with eighteen (18), sixteen (16), and thirty-two (32) calibration profiles, leaving five (5), three (3), and seven (7) degrees of freedom for error, and computing performance with sixteen (16), four (4), and twelve (12) holdout profiles. For all three data sets and varying conditions, the study failed to replicate superiority of Hagerty's segment-level conjoint method. Lack of degrees of freedom, it seems, is not enough to explain individual-level conjoint models' superiority. However, it seems that performance measures, and relative performance of model forms, are influenced by the number of holdout profiles used as the basis for comparison. No study has been performed yet that could shed light on this speculation, as it is hard to believe that those empirical studies' findings are just "a fluke" (Green, Krieger, and Schaffer 1993a, p. 346).

One major area for concern is a lack of tests for segment-level performance measures. Testing procedures that may be used for individual-level estimates are mostly not possible due to insufficient data, or they are not valid due to violation of test assumptions. Reibstein, Bateson, and Boulding's (1988) choice to use the alpha level of a Chow test as a measure of reliability has already been exposed as inappropriate (Green and Srinivasan 1990, p. 12).

The issue of conflictual effects of increased individual variance versus decreased bias in parameter estimates, and vice versa, is surfacing at several locations in this study: When deciding upon the appropriate conjoint model form, inclusion of interaction terms increases the number of parameters in the model, reducing the number of error degrees of freedom, thus increasing bias in estimation. Estimating segment-level conjoint models increases the individual variances, but also increases the degrees of freedom for error, thus decreasing bias in estimation, and making all parameters for the segment-level models significant. In this study, bias of parameter estimates in the individual-level models does not seem to be high enough to outweigh increases in individual variances for the segment-level models. Therefore, there is no gain in accuracy of prediction with between-subjects conjoint models.

Maybe one of the most important findings in this study concerns appropriateness of between-subjects standardization or normalization of conjoint part-worths before application of clustering or other segment-level aggregation methods: These procedures change the relative impact of attribute levels, and subsequently the relative importance of attributes on overall product utility. Whatever clusters existed before application of such procedures, they are destroyed afterwards. Therefore, and as has been demonstrated in section 4.3.1 on pp. 198, it is important to apply appropriate scaling to the original regression coefficients depending on what insights one expects from further examination of part-worth utilities, or how one intends to utilize them in subsequent procedures. Though standardization within subjects does not change within-subject attribute importances, differences in "intensity" of ratings across subjects are lost. There is nothing known about possible consequences of these data manipulations on subsequent results of clustering procedures.

Simply applying an index of clusterability to the original regression coefficients, mistakenly termed part-worths, with the consequence of "implying the non-clusterability of the respondents" (Akaah and Korgaonkar 1988, p. 41) is just as inappropriate as statements like those of Hagerty (1985, p. 170): "... these types of clustering retain the idea that clusters *exist* ... On the contrary, the plots of actual respondents show no obvious clusters at all. Therefore, why should we not do away with the idea of clusters completely ?" In contrast to those and similar statements in the literature, clustering of part-worth utilities in the current study yielded valuable information about market segments and their profiles based on benefit attributions to product features, though these clusters were not *obvious* but opaque. These findings allow development of products that appeal to specific market segments as well as adjustment of communication targeted to selected segments.

Different clustering procedures, however, show varying ability to group respondents into meaningful subdivisions for target marketing. Fuzzy clustering performed best and worst for all cluster methods in terms of prediction, and in terms of substantive interpretation of cluster profiles, depending on the degree of fuzziness allowed. While the improvement with fuzzy clustering over hard clustering methods is encouraging and should be explored further in future studies, it was not enough to reach predictive accuracy of individual-level conjoint models.

5.2 Contribution of Systems Science to This Study

Systems thinking invisibly influenced this study at two levels of the inquiry process:

(1) at the level of the topic or subject area, as this study examines measurements of customer value systems (micro view), and

(2) at the level of the inquiry process itself, i.e. in the way of analyzing and approaching the problem which reveals to be the more important contribution (macro view).

Conceptually, this study may be regarded as an exercise in system identification, i.e. identification of the customer value system, for a notebook computer with the methodology of conjoint analysis, as well as an exercise in possibilities for improvement in the identification process with selected methodological variations. This study examined if a system of conjectured decision criteria (i.e. attributes as elements of the value system) is an accurate representation of customer value structure, i.e. of a customer´s value system. Specific aspects of different measurements and representations were tested. These tests concern influences of variations in the conjoint method on attributions of benefits to attribute levels, i.e. to the elements of the system. At the same time, this study allows statements about the relationship between elements of the customer value system, for instance if these relationships can be represented as a set of simple algebraic rules. Questions that could be answered after system identification are, for instance, questions like: Is this understanding of the customer value system able to predict behavior, i.e. system outcome, and to what degree, or, is it possible to identify the customer value system with respective estimation methods better than with direct questions about benefit attributions to product attributes ? Believabiliy of selected research questions was mainly tested with performance measures of system outcome.

While those questions represent an important part of systems thinking in this study, it is the approach, the perspective taken where systems methodology came to bear most, as, for instance, in decisions about the scope of the dissertation: The decision to expand on the performance measures included in the study is a result of the belief that judgments about the relative pērformance of customer value systems are only adequate when they include different aspects of the systems´ behaviors. Statements about the systems´ behaviors and relative performance of different representations and/or different measurements are only valid when purpose and context are clearly defined. If these two characteristics, or conceptual companions of a system are not clearly defined, statements about relative performance are meaningless. As performance measures differ in their ability to represent different purposes and contexts it seemed appropriate to calculate several different performance measures, thus allowing for a much more comprehensive understanding of effects of methodological variations on system performance.

Another aspect where systems thinking comes to the fore, is in the belief that there are no universal criteria for comparisons, i.e. compromises are inevitable: Criteria for measurement methods, like comparability, optimality, generalizability, or objectives and their achievement, have different repercussions on precision, certainty, and usefulness of respective results of measurements. Therefore, as results of measurements are dependent on measurement conditions, interpretation of results seems only possible with clear definition of purpose and context of the system studied. This thinking qualifies, or even diminishes belief in tests of statistical significance, and this thinking is supported with results of Monté Carlo studies, for instance by Umesh and Mishra (1990), which gauge dependencies of measurement conditions on results.

As a final example, systems thinking may be responsible for the detection of a conceptual fallacy concerning segmentation with cluster algorithms in connection with conjoint methodology: Standardization of part-worth utilities obtained with conjoint methods before the application of cluster algorithms changes the cluster object, i.e. the decision context, and is therefore not appropriate (cp. p. 198). A different measure for between-subject comparisons of part-worth utilities is necessary (scaled part-worths). Application of systems thinking revealed a violation of dependencies that may have gone undetected with a different conceptual approach.

5.3 Contribution to Marketing Theory and Practice

This study contributes to marketing theory in four areas where methodological problems have been identified for conjoint analysis in the literature (cp. section 2.4 on page 49):

Influences of the type of attribute, specifically of solely technical or product-referent and user-referent attributes, on prediction and resulting value structure has been examined. The type of attribute has the ability to significantly influence accuracy of prediction. However, it did not significantly influence value structure, i.e. the relative importance of different attributes. The usefulness of inclusion of user-referent attributes cannot be stated in general terms. In this study, a marketer could make use of positive effects of firm reputation in form of adapted product offerings and communication policy, but for a different product, and a different type of user-referent attribute this need not be the case. However, this study´s results should encourage more detailed and more extensive studies of effects of different types of attributes on prediction and value structure of potential customers. Conjectured problems with different fractional factorial designs could not be substantiated. Quite to the contrary, properly derived fractional factorial designs had no noticeable distorting effects on customer value. Thus, this study contributed to the notion of conjoint analysis as a reliable method for measurement of customer value.

This study also confirmed superiority of individual-level conjoint models over segment-level models, contributing to the scarce literature of empirical studies testing suggested improvements in prediction with aggregate methods. At the same time, possible reasons for failure of this empirical study to replicate theoretical findings have been exposed but need further study. One of the keys to this understanding seem to be in the performance measures used, and in the bases for their computation. Too little is known about those measures´ properties to allow for conclusions.

In addition, this research confirmed that main-effects-only models may still be superior to models with interactions, even with the presence of non-metric and non-monotone attribute levels. Extending methodological variations to models with interaction terms showed no significant gains for prediction but also no problematic distortions of value structure. Nevertheless, though the researcher may rest confident that main-effects-only models perform very well in most cases, he should reserve the possibility to check for the necessity to include them in the conjoint model with appropriate precautions in the design of the stimulus sets. It is usually not possible to estimate models with interaction terms when their inclusion has not been taken into account in the design phase of the study.

This study also revealed that repeated measures may provide valuable information about the relative influences of treatments and individual variation. Without repeated measures, treatment effects are only revealed when they are much larger than individual variation. With knowledge of individual variation, it is possible to gauge which absolute magnitudes in the effects should be considered substantively meaningful, and from which magnitudes of changes on. Once the magnitudes of effects of individual methodological variations is known it is possible to combine those variations in conjoint studies that are likely to cancel out in their effects on predictive performance and measurement of value structure. This allows for increased reliability of the measurement instrument.

This study provides for the first application of Chiu´s 'subtractive clustering' algorithm (1994) in marketing in order to determine the adequate number of clusters for fuzzy clustering, but without success. Application of this procedure to a normalized data set of part-worth utilities is inappropriate as it distorts within-subject part-worth utilities. Non-normalized application of the algorithm leaves the researcher within a territory of unknown theoretical properties, and it did not yield valid cluster solutions. Another important contribution of this study is the exposure of selected scaling procedures as inappropriate for clustering purposes in connection with conjoint measurement. This finding, as well as easy availability of computer programs, emphasizes the necessity to carefully examine the presence or absence of statistical and computational assumptions, as validity of the findings hinges on the proper application of methods.

The current study provides some support for the scepticism against findings of early conjoint studies in the literature, and also of commercial conjoint studies reported today that violate some statistical assumptions (e.g. no holdout judgments in 91% of commercial studies in Europe; Wittink, Vriens, and Burhenne 1994, p. 47). Careful examination of this study´s data and associated test assumptions suggests that results of conjoint studies using only one performance measure cannot be relied on, as different measures may perform very differently in tests. Consistent results of tests over a variety of performance measures, however, may increase the belief in general statements about conjoint model behavior. This study was conducted with a variety of performance measures, and under varying measurement conditions, increasing belief in the study´s findings.

Second, some ratings-based and choice-based performance measures, especially Fisher's z(r_{xy}) and First-Hit, revealed the need for different test procedures as specifically First-Hit often seriously violated the assumption of normal distribution of the data that is a prerequisite for validity of most parametric tests. However, for reasons of comparability, the same test procedures were applied to all performance measures, sometimes with divergent results. In this study, such differences could be resolved with examinations of respective distributional assumptions of the data (normality) and/or other prerequisites of the tests (e.g. concerning the number of data points in the case of Chi-Square, or the absence of an interval scale for r_{xy}), allowing for valid conclusions. With these procedures, this research study provides greater confidence in its findings than is possible for some prior studies. Related to this issue of divergent results is the question of the appropriate performance measure for conjoint experiments: ratings-based or choice-based measures, especially r_{xy} or First-Hit. It seems that part of the dissension in the literature could be resolved with careful examination of presence or absence of distributional assumptions of the data and/or prerequisites of testing procedures.

5.4 Limitations

Main limitations of this study that could be determined in the design phase have already been exposed in the introduction to this research. Some further limitations have surfaced, since, or should be mentioned for completeness.

In the narrow sense, findings of this study are only generalizable to the immediate research conditions, for instance the product class under review, a notebook computer, i.e. a product category that is relatively new and moderately complex. However, together with studies that used more familiar products, like apartments or yoghurt, and which yielded similar results of high reliability for conjoint analysis, the current study's findings may be generalized to a much broader class of products that includes relatively new product categories, i.e. products that combine attributes in a new manner, or provide benefits not possible with current market offerings.

Furthermore, most conjoint studies employ descriptions of laboratory or experimental products which is a necessity of fractional factorial designs. However, it is not known in how far conjoint studies could benefit from actual products for evaluation. The construction of Pareto-optimal sets of stimuli has already been demonstrated to limit the number of comparisons necessary (Krieger and Green 1991), but evaluation and choices of holdouts or other surrogate procedures for market choice cannot replace peformance evaluation on the basis of actual purchases, which this study also fails to be able to conduct. Greater nearness to actual choices is still highly desirable.

Finally, choice of a student population may have helped in raising the level of accuracy in prediction. However, for the study's kind of findings, i.e. the influence of methodological variations on performance and reliability, the absolute level of performance reached is not of primary relevance. Instead, it is the relative effects exposed, and conjoint measurement's insensitivity towards them that is of primary concern. Thus, taking a student population does not limit this study's findings.

5.5 Directions for Future Research

Concerning this study's empirical finding that segment-level conjoint models do not increase accuracy in prediction of product preference and choice behavior in comparison to individual-level conjoint models, and taking into account the limited evidence of other studies conducted to date, three (3) ways to increase accuracy in measurement and associated prediction seem viable:

(1) Increasing individualization of the preference / choice task;

(2) Repeated measurements;

(3) Gathering additional information about respondents.

Cf. (1): Boecker and Schweikl´s study (1988) is still the only one to have attempted this approach. One may speculate, due to the unavailability of their computer program, and the immense effort to develop one of one´s own, this remains the only application to date, though individualization of attributes seems a viable way to increase accuracy of conjoint experiments. However, managerial relevance of this approach may be doubted, as appeal of a limited set of product attributes to a great number of potential market participants is of greater concern in practice.

Cf. (2): In all those cases of conjoint models where time was the only variation between two measurements, the second (i.e. repeated) measurement led to increased accuracy in prediction. Though one must be cautious about a possible learning effect, a repeated measurement promises to increase accuracy more than filigree work with respect to intricacies of estimation method and further methodological developments.

Cf. (3): Another promising but rather costly approach to value measurement has just recently been demonstrated by Sukhdial, Chakraborty, and Steger (1995), combining information about social values of respondents with value attributions to product profiles, e.g. LOV-scale and conjoint measurement of luxury cars, thereby increasing overall accuracy in prediction of car ownership, but also increasing usefulness for adaptation of communication policy. Such combinations may expose greater potential for increases in prediction than further exploration of conjoint variations.

An important gauge to judge attempts for further improvements in measurement accuracy is the respondent itself, and two related questions:

(1) How accurately can people be measured ? Where are possible limits of accuracy in measurement of people ?

(2) How can we improve estimation of the distinctness of preferences or choices ?

In order to answer the first question, we need many more "roadmaps" as guides for the choice of conjoint methodological variations similar to those we can take for granted in other areas of statistical methodology, as for instance in regression: There, we know properties of methods and consequences from violations of assumptions much better than for conjoint analysis. The second question is also difficult to answer, but it seems that repeated measurements would be a viable approach to elicit stability of preferences.

Application of fuzzy logic as a concept to address inherent uncertainty in the measurement object also seems to be a viable approach, and it would be helpful to know more about the method´s relative superiority vs. deterministic and statistical models. It also seems necessary to conduct more studies with actual choices as the basis for performance measures, realizing that in many instances this would be too expensive.

For practical application of conjoint studies, current programs do not support the researcher very well. SPSS´ Categories program, for instance, can only address main effects models, and does not support controlled development of designs. The researcher has to develop a lot of analytical tools himself, limiting the ability to apply a great number of methods in commercial studies. Even more dangerous and detrimental to conjoint measurement´s reputation as a good tool for customer value measurement is application of programs with limited flexibility, tempting commercial researchers to take unsupported shortcuts.

This study´s overviews and tests with selected performance measures for evaluation of accuracy in prediction underscores the urgent need to better understand properties of different performance measures under varying conditions of conjoint measurements. Current Monté Carlo studies lack in breadth of parameters included, and in depth of parameter ranges which limits their usefulness for interpretation of current studies. As for Monté Carlo studies for segment-level conjoint models, it seems premature to conduct them before important conceptual problems have been resolved, as for instance how to adequately test differences among segment-level models. Nevertheless, apart from theoretical studies about performance measures, Monté Carlo studies suggest a viable way to expose properties of performance measures under varying methodological conditions.

REFERENCES

Addelman, Sidney (1962) "Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments," Technometrics, Vol. 4, n. 1, February, pp. 21 - 58

Ajzen, Icek and Martin Fishbein (1980 Understanding Attitudes and Predicting Social Change, Englewood Cliffs, NJ: Prentice-Hall

Akaah, Ishmael P. (1988) "Cluster Analysis Versus Q-Type Factor Analysis as a Disaggregation Method in Hybrid Conjoint Modeling: An Empirical Investigation," Journal of the Academy of Marketing Science, Vol. 16, n. 2, Summer, pp. 11 - 18

Akaah, Ishmael P. (1991) "Predictive Performance of Self-Explicated, Traditional Conjoint, and Hybrid Conjoint Models under Alternative Data Collection Modes," Journal of the Academy of Marketing Science, Vol. 19, n. 4, Fall, pp. 309 - 314

Akaah, Ishmael P. and Pradeep K. Korgaonkar (1983) "An Empirical Comparison of the Predictive Validity of Self-Explicated, Hüber-Hybrid, Traditional Conjoint, and Hybrid Conjoint Models," Journal of Marketing Research, Vol. XX, n. 2, May, pp. 187 - 197

Alba, Joseph W. and J. Wesley Hutchinson (1987) "Dimensions of Consumer Expertise," Journal of Consumer Research, Vol. 13, n. 4, March, pp. 411 - 454

Anderson, Norman H. (1981) Foundations of Information Integration Theory, New York: Academic Press

Anderson, Norman H. (1982) Methods of Information Integration Theory, New York: Academic Press

Bagozzi, Richard P. (1982) "A Field Investigation of Causal Relations Among Cognitions, Affect, Intentions, and Behavior," Journal of Marketing Research, Vol. XIX, n. 4, November, pp. 562 - 584

Bateson, John E.G., David Reibstein, and William Boulding (1987) "Conjoint Analysis Reliability and Validity: A Framework for Future Research," in: Houston, Michael J. (Ed.), Review of Marketing, Chicago, IL: AMA, pp. 451 - 481

Bettman, James R., Noël Capon, and Richard J. Lutz (1975) "Cognitive Algebra in Multi-Attribute Attitude Models," Journal of Marketing Research, Vol. XII, n. 2, May, pp. 151 - 164

Bezdek, James C. (1981) Pattern Recognition With Fuzzy Objective Function Algorithm, New York: Plenum Press

Boecker, Franz and Herbert Schweikl (1988) "Better Preference Prediction With Individualized Sets of Relevant Attributes," International Journal of Research in Marketing, Vol. 5, n. 1, September, pp. 15 - 24

Burnkrant, Robert E. and Thomas J. Page, Jr. (1982) "An Examination of the Convergent, Discriminant, and Predictive Validity of Fishbein's Behavioral Intention Model," Journal of Marketing Research, Vol. XIX, n. 4, November, pp. 550 - 561

Campbell, Donald T. and Donald W. Fiske (1959) "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," Psychological Bulletin, Vol. 56, n. 2, March, pp. 81 - 105

Campbell, John P. (1976) "Psychometric Theory," in: Dunnette, Marvin D. (Ed.), Handbook of Industrial and Organizational Psychology, Ch. 6, Chicago, IL: Rand McNally College Publishing, pp. 185 - 222

Cattin, Philippe and Dick R. Wittink (1982) "Commercial Use of Conjoint: A Survey," Journal of Marketing, Vol. 46, Summer, pp. 44 - 53

Chiu, S. (1994) "Fuzzy Model Identification Based on Cluster Estimation," Journal of Intelligent & Fuzzy Systems, Vol. 2, n. 3, September, pp. 267 - 278

Christopher, Martin (1982) "Value-In-Use Pricing," European Journal of Marketing, Vol. 16, n. 5, pp. 35 - 46

Chuah, Siew and James C. Bezdek (1987) "Optimal Classifier Design Using Fuzzy k-Nearest Neighbor Rules," in: Kacprzyk, J. and S. A. Orlovski (Eds.), Optimization Models Using Fuzzy Sets and Possibility Theory, Boston, MA: D. Reidel Publishing, pp. 432 - 446

Churchill, Jr., Gilbert A. (1979) "A Paradigm for Developing Better Measures of Marketing Constructs," Journal of Marketing Research, Vol. XVI, n. 1, February, pp. 64 - 73

Cohen, Joel B. (1979) "The Structure of Product Attributes: Defining Attribute Dimensions for Planning and Evaluation," in: A. Shocker (Ed.), Analytic Approaches to Product and Marketing Planning, Cambridge, MA: Marketing Science Institute

Connor, William S. and Shirley Young (1961) Fractional Factorial Designs for Experiments With Factors at Two and Three Levels, US Department of Commerce, National Bureau of Standards, Applied Mathematics Series, No. 58, Washington, DC: US Government Printing Office, September 1

Corfman, Kim P. (1991a) "Comparability and Comparison Levels Used in Choices Among Consumer Products," Journal of Marketing Research, Vol. XXVIII, n. 3, August, pp. 368 - 374

Darmon, René Y. and Dominique Rouziès (1989) "Assessing Conjoint Analysis Internal Validity: The Effect of Various Continuous Attribute Level Spacings," International Journal of Research in Marketing, Vol. 6, n. 1, September, pp. 35 - 44

Darmon, René Y. and Dominique Rouziès (1991) "Internal Validity Assessment of Conjoint Estimated Attribute Importance Weights," Journal of the Academy of Marketing Science, Vol. 19, n. 4, Fall, pp. 315 - 322

DeSarbo, Wayne S., Michel Wedel, Marco Vriens, and Venkatram Ramaswamy (1992) "Latent Class Metric Conjoint Analysis," Marketing Letters, Vol. 3, n. 3, August, pp. 273 - 288

DeSarbo, Wayne S., Richard L. Oliver, and Arvind Rangaswamy (1989) "A Simulated Annealing Methodology for Clusterwise Linear Regression," Psychometrica, Vol. 54, n. 4, December, pp. 707 - 736

Dyer, James S.; Saaty, Thomas L.; Harker, Patrick T. and Luis G. Vargas; Dyer, James S. (1990) "Remarks on the Analytic Hierarchy Process; An Exposition of the AHP in Reply to the Paper 'Remarks on the Analytic Hierarchy Process'; Reply to 'Remarks on the Analytic Hierarchy Process' by J.S. Dyer; A Clarification of 'Remarks on the Analytic Hierarchy Process'," Management Science, Vol. 36, n. 3, March, pp. 249 - 275

Elrod, Terry, Jordan J. Louviere, and Krishnakumar S. Davey (1992) "An Empirical Comparison of Ratings-Based and Choice-Based Conjoint Models," Journal of Marketing Research, Vol. XXIX, n. 3, August, pp. 368 - 377

Engel, James F., Roger D. Blackwell, and Paul W. Miniard (1990) Consumer Behavior, 6th Ed., Chicago: Dryden Press

Fishbein, Martin (1963) "An Investigation of Relationship between Beliefs about an Object and the Attitude toward the Object," Human Relations, Vol. 16, pp. 233 - 240

Fishbein, Martin and Icek Ajzen (1975) Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research, Reading, MA: Addison-Wesley

Forbis, John L. and Nitin T. Mehta (1981) "Value-Based Strategies for Industrial Products," Business Horizons, Vol. 24, n. 3, May-June, pp. 32 - 42

Geistfeld, L. V., G. B. Sproles, and S. B. Badenhop (1977) "The Concept and Measurement of a Hierarchy of Product Characteristics," in: W. D. Perreault, Jr. (Ed.), Advances in Consumer Research, Vol. 4, Atlanta: Association for Consumer Research, pp. 302 - 307

Gensch, Dennis H. and Sanjoy Ghose (1992) "A Dimensional versus Attribute Approach For Disaggregate Choice Models," Marketing Letters, Vol. 3, n. 1, February, pp. 27 - 37

Green, P. E. and V. R. Rao (1971) "Conjoint Measurement for Quantifying Judgmental Data," Journal of Marketing Research, Vol. VIII, n. 3, August, pp. 355 - 363

Green, Paul E. and Abba M. Krieger (1991) "Segmenting Markets With Conjoint Analysis," Journal of Marketing, Vol. 55, n. 4, October, pp. 20 - 31

Green, Paul E. and Abba M. Krieger (1993) "Conjoint Analysis with Product-Positioning," in: Eliashberg, Jehoshua and Gary L. Lilien (Eds.), Marketing: Handbooks in OR and MS, Vol. 5, New York: North-Holland, Ch. 10, pp. 467 - 515

Green, Paul E. and Catherine M. Schaffer (1991) "Importance Weight Effects On Self-Explicated Preference Models: Some Empirical Findings," in: Holman, Rebecca M. and Michael R. Solomon (Eds.), Advances in Consumer Research, Vol. 18, Provo, UT: Association of Consumer Research, pp. 476 - 482

Green, Paul E. and Kristiaan Helsen (1989) "Cross-Validation Assessment of Alternatives to Individual-Level Conjoint Analysis: A Case Study," Journal of Marketing Research, Vol. XXVI, n. 3, August, pp. 346 - 350

Green, Paul E. and V. Srinivasan (1978) "Conjoint Analysis in Consumer Research: Issues and Outlook," Journal of Consumer Research, Vol. 5, n. 2, September, pp. 103 - 123

Green, Paul E. and V. Srinivasan (1990) "Conjoint Analysis in Marketing: New Developments With Implications for Research and Practice," Journal of Marketing, Vol. 54, n. 4, October, pp. 3 - 19

Green, Paul E. and Yoram Wind (1975) "New Way to Measure Consumers´ Judgments," Harvard Business Review, Vol. 53, July-August, pp. 107 - 117

Green, Paul E., Abba M. Krieger, and Catherine M. Schaffer (1993a) "An Empirical Test of Optimal Respondent Weighting in Conjoint Analysis," Journal of the Academy of Marketing Science, Vol. 21, n. 4, Fall, pp. 345 - 351

Green, Paul E., Abba M. Krieger, and Catherine M. Schaffer (1993b) "A Hybrid Conjoint Model With Individual-Level Interaction Estimation," in: McAlister, Leigh and Michael L. Rothschild (Eds.), Advances in Consumer Research, Vol. XX, Provo, UT: Association for Consumer Research, pp. 149 - 154

Green, Paul E., Abba M. Krieger, and P. Bansal (1988) "Completely Unacceptable Levels in Conjoint Analysis: A Cautionary Note," Journal of Marketing Research, Vol. XXV, n. 3, August, pp. 293 - 300

Green, Paul E., J. Douglas Carroll, and Frank J. Carmone (1978) "Some New Types of Fractional Factorial Designs for Marketing Experiments," Research in Marketing, Vol. 1, pp. 99 - 122

Green, Paul E., Kristiaan Helsen, and Bruce Shandler (1988) "Conjoint Internal Validity Under Alternative Profile Presentations," Journal of Consumer Research, Vol. 15, n. 3, December, pp. 392 - 397

Haas, Robert W. (1989) Industrial Marketing Management, 4th Ed., Boston, MA: PWS-Kent Publishing

Hagerty, Michael R. (1985) "Improving the Predictive Power of Conjoint Analysis: The Use of Factor Analysis and Cluster Analysis," Journal of Marketing Research, Vol. XXII, n. 2, May, pp. 168 - 184

Hagerty, Michael R. (1986) "The Cost of Simplifying Preference Models," Marketing Science, Vol. 5, n. 4, Fall, pp. 298 - 319

Hagerty, Michael R. (1993) "Can Segmentation Improve Predictive Accuracy in Conjoint Analysis ?," Journal of the Academy of Marketing Science, Vol. 21, n. 4, Fall, pp. 353 - 355

Hagerty, Michael R. and V. Srinivasan (1991) "Comparing the Predictive Powers of Alternative Multiple Regression Models," Psychometrica, Vol. 56, n. 1, March, pp. 77 - 85

Hair, Joseph F., Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black (1992) Multivariate Data Analysis with Readings, 3rd Ed., New York: Macmillan Publishing

Hauser, John R. and Glen L. Urban (1986) "The Value Priority Hypotheses for Consumer Budget Plans," Journal of Consumer Research, Vol. 12, n. 4, March, pp. 446 - 462

Helsen, Kristiaan and Paul E. Green (1993) "A Computational Study of Replicated Clustering With An Application to Marketing Segmentation," American Marketing Association Winter Educators' Conference, Vol. 4, p. 329

Hines, William W. and Douglas C. Montgomery (1990) Probability and Statistics in Engineering and Management Science, 3rd Ed., New York: John Wiley & Sons

Hosseini, Jamshid C. (1987) Segment Congruence Analysis: An Information-Theoretic Approach, Unpublished Doctoral Dissertation, Portland State University

Hruschka, H. (1986) "Market Definition and Segmentation Using Fuzzy Clustering Methods," International Journal of Research in Marketing, Vol. 3, n. 2, pp. 117 - 134

Jain, Arun K., Franklin Acito, Naresh K. Malhotra, and Vijay Mahajan (1979) "A Comparison of the Internal Validity of Alternative Parameter Estimation Methods in Decompositional Multiattribute Preference Models," Journal of Marketing Research, Vol. XVI, n. 3, August, pp. 313 - 322

Johnson, Eric J., Robert J. Meyer, and Sanjoy Ghose (1989) "When Choice Models Fail: Compensatory Models in Negatively Correlated Environments," Journal of Marketing Research, Vol. XXVI, n. 3, August, pp. 255 - 270

Johnston, J. (1984) Econometric Models, 3rd Ed., New York: McGraw-Hill

Kamakura, Wagner A. (1988) "A Least Squares Procedure for Benefit Segmentation with Conjoint Experiments," Journal of Marketing Research, Vol. XXV, n. 2, May, pp. 157 - 167

Kaufman, Leonard and Peter J. Rousseeuw (1990) Finding Groups in Data, New York: John Wiley & Sons

Kosko, Bart (1993) Fuzzy Thinking: The New Science of Fuzzy Logic, New York: Hyperion

Kotler, Philip (1991) Marketing Management, 7th Ed., Englewood Cliffs, NJ: Prentice-Hall

Krieger, Abba M. and Paul E. Green (1991) "Designing Pareto Optimal Stimuli for Multiattribute Choice Experiments," Marketing Letters, Vol. 2, n. 4, November, pp. 337 - 348

Kuhn, Thomas (1970) The Structure of Scientific Revolutions, 2nd Ed., Chicago: The University of Chicago Press

Laitamaki, Jukka M. and Leo M. Renaghan (1988) "Value Based Pricing Strategies for Services: An Empirical Study of the Hotel Industry," in: Carol Surprenant (Ed.) Add Value to your Service, Chicago: American Marketing Association, pp. 179 - 183

Lee, Yvonne (1994) "AT&T, IBM forge ahead in PDA market despite low demand," InfoWorld, August 1, p. 6

Leigh, Thomas W., David B. MacKay, and John O. Summers (1981) "On Alternative Experimental Methods for Conjoint Analysis," in: Monroe, Kent B. (Ed.), Advances in Consumer Research, Vol. VIII, Urbana, IL: Association for Consumer Research, pp. 317 - 322

Leigh, Thomas W., David B. MacKay, and John O. Summers (1984) "Reliability and Validity of Conjoint Analysis and Self-Explicated Weights: A Comparison," Journal of Marketing Research, Vol. XXI, n. 4, November, pp. 456 - 462

Lendaris, George G. (1986) "On Systemness and the ProblemSolver: Tutorial Comments," IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-16, n. 4, July/August, pp. 604 - 610

Lilien, Gary L., Philip Kotler, and K. Sridhhar Moorthy (1992) Marketing Models, Englewood Cliffs, NJ: Prentice-Hall

Louviere, J. J. (1974) "Predicting the Evaluation of Real Stimulus Objects from an Abstract Evaluation of their Attributes: The Case of Trout Streams," Journal of Applied Psychology, Vol. 59, n. 5, pp. 572 - 577

Louviere, Jordan L. (1988) Analyzing Decision Making: Metric Conjoint Analysis, Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 67, Beverly Hills, CA: Sage Publications

McCullough, James and Roger Best (1979) "Conjoint Measurement: Temporal Stability and Structural Reliability," Journal of Marketing Research, Vol. XVI, n. 1, February, pp. 26 - 31

Mehrotra, Sunil and John Palmer (1985) "Relating Product Features to Perceptions of Quality: Appliances," in: J. Jacoby and J. Olson (Eds.), Perceived Quality, Ch. 5, Lexington, MA: Lexington Books, pp. 81 - 96

Mehta, Raj, William L. Moore, and Teresa M. Pavia (1992) "An Examination of the Use of Unacceptable Levels in Conjoint Analysis," Journal of Consumer Research, Vol. 19, n. 3, December, pp. 470 - 476

Miller, G. A. (1956) "The Magic Number Seven, Plus or Minus Two: Some Limits on Our Capacity For Processing Information," Psychological Review, Vol. 63, n. 2, March, pp. 81 - 97

Monroe, Kent B., Akshay R. Rao, and Joseph D. Chapman (1987) "Toward a Theory of New-Product Pricing," in Frazier, Gary L. and Jagdish N. Sheth (Eds.), Contemporary Views on Marketing Practice, Chapter 10, pp. 201 - 213

Montgomery, David B. (1986) "Conjoint Calibration of the Customer/Competitor Interface in Industrial Markets," in: Backhaus, Klaus and David T. Wilson (Eds.), Industrial Marketing: A German-American Perspective, Berlin: Springer-Verlag, pp. 297 - 319

Moore, David S. and George P. McCabe (1989) Introduction to the Practice of Statistics, New York: W.H. Freeman

Moore, William L. and Morris B. Holbrook (1990) "Conjoint Analysis on Objects with Environmentally Correlated Attributes: The Questionable Importance of Representative Design," Journal of Consumer Research, Vol. 16, n. 4, March, pp. 490 - 497

Moore, William L. and Richard J. Semenik (1988) "Measuring Preferences with Hybrid Conjoint Analysis: The Impact of a Different Number of Attributes in the Master Design," Journal of Business Research, Vol. 16, pp. 261 - 274

Myers, James H. and Allan D. Shocker (1981) "The Nature of Product-Related Attributes," Research in Marketing, Vol. 5, pp. 211 - 236

Naik, Gautam (1994) "AT&T Says EO Unit Will Close as Sales Of 'Personal Data Communicators' Lag," The Wall Street Journal, Thursday, July 28, p. B5

Nataraajan, Rajan (1993) "Prediction of Choice in a Technically Complex, Essentially Intangible, Highly Experiential, and Rapidly Evolving Consumer Product," Psychology and Marketing, Vol. 10, n. 5, September/October, pp. 367 - 379

Norusis, Marija J./SPSS Inc. (1993) SPSS Base System User´s Guide, Release 6.0, Chicago, IL: SPSS, Inc.

Ogawa, Kohsuke (1987) "An Approach to Simultaneous Estimation and Segmentation in Conjoint Analysis," Marketing Science, Vol. 6, n. 1, pp. 66 - 81

Olson, Jerry C. and Thomas J. Reynolds (1983) "Understanding Consumers´ Cognitive Structures: Implications for Advertising Strategy," in: L. Percy and A. Woodside (Eds.), Advertising and Consumer Psychology, Lexington, MA: Lexington Book

Oppewal, Harmen, Jordan J. Louviere, and Harry J.P. Timmermans (1994) "Modeling Hierarchical Conjoint Processes With Integrated Choice Experiments," Journal of Marketing Research, Vol. XXXI, n. 1, February, pp. 92 - 105

Ostrom, Amy and Dawn Iacobucci (1995) "Consumer Trade-Offs and the Evaluation of Services," Journal of Marketing, Vol. 59, n. 1, January, pp. 17 - 28

Pedhazur, Elazar J. (1982) Multiple Regression in Behavioral Research: Explanation and Prediction, Fort Worth, TX: Harcourt Brace College Publishers

Peter, J. Paul (1979) "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," Journal of Marketing Research, Vol. XVI, n. 1, February, pp. 6 - 17

Potter, Donald V. (1988) "The Two Best Consultants in the World," Business Horizons, Vol. 31, n. 5, September/October, pp. 25 - 28

Reibstein, David, John E. G. Bateson, and William Boulding (1988) "Conjoint Analysis Reliability: Empirical Findings," Marketing Science, Vol. 7, n. 3, Summer, pp. 271 - 286

Rokeach, M. J. (1973) The Nature of Human Values, New York: The Free Press

Ruspini, Enrique H. (1970) "Numerical Methods for Fuzzy Clustering," Information Science, Vol. 2, pp. 319 - 350

Saaty, Thomas L. (1978) "Exploring the Interface Between Hierarchies, Multiple Objectives and Fuzzy Sets," Fuzzy Sets and Systems, Vol. 1, n. 1, pp. 57 - 68

Saaty, Thomas L. (1980) The Analytic Hierarchy Process, New York: McGraw-Hill

SAS Institute (1994a) JMP Statistics and Graphics Guide, Version 3, Cary, NC: SAS Institute Inc.

SAS Institute (1994b) JMP Changes and Enhancements, Version 3.1, Cary, NC: SAS Institute Inc.

Sheth, Jagdish N., Bruce I. Newman, and Barbara L. Gross (1991a) "Why We Buy What We Buy: A Theory of Consumption Values," Journal of Business Research, Vol. 22, pp. 159 - 170

Sheth, Jagdish N., Bruce I. Newman, and Barbara L. Gross (1991b) Consumption Values and Market Choices, Cincinnati, OH: South-Western Publishing

Slovic, P. and S. Lichtenstein (1968) "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment," Organizational Behavior and Human Performance, Vol. 6, pp. 649 - 744

Slovic, P., B. Fischhoff, and S. J. Lichtenstein (1977) "Behavioral Decision Theory," Annual Review of Psychology, Vol. 28, pp. 1 - 39

SPSS, Inc. (1994) SPSS 6.1 Categories®, Version 6.1, Chicago, IL: SPSS, Inc.

Srinivasan, V. (1988) "A Conjunctive-Compensatory Approach to the Self-Explication of Multiattributed Preferences," Decision Sciences, Vol. 19, n. 2, Spring, pp. 295 - 305

Srinivasan, V. and Gordon A. Wyner (1989) "CASEMAP: Computer-Assisted Self-Explication of Multiattributed Preferences," in: Henry, Walter, Michael Menasco, and Hirokazu Takada, New Product Development and Testing, Lexington, MA: Lexington Books, pp. 91 - 111

Sriram, Ven and Andrew M. Forman (1993) "The Relative Importance of Products´ Environmental Attributes: A Cross-cultural Comparison," International Marketing Review, Vol. 10, n. 3, pp. 51 - 70

Steckel, Joel H., Wayne S. DeSarbo, and Vijay Mahajan (1991) "On the Creation of Acceptable Conjoint Analysis Experimental Designs," Decision Sciences, Vol. 22, pp. 435 - 442

Steenkamp, Jan-Benedict E.M. and Dick R. Wittink (1994) "The Metric Quality of Full-Profile Judgments and the Number-of-Attribute-Levels Effect in Conjoint Analysis," International Journal of Research in Marketing, Vol. 11, n. 3, pp. 275 - 286

Sukhdial, Ajay S., Goutam Chakraborty, and Eric K. Steger (1995) "Measuring Values Can Sharpen Segmentation in the Luxury Auto Market," Journal of Advertising Research, Vol. 35, n. 1, January/February, pp. 9 - 22

Umesh, U. N. and Sanjay Mishra (1990) "A Monté Carlo Investigation of Conjoint Analysis Index-of-Fit: Goodness of Fit, Significance, and Power," Psychometrica, Vol. 55, n. 1, March, pp. 33 - 144

Urban, Glenn L, John R. Hauser, and Nikhilesh Dholakia (1987) Essentials of New Product Management, Englewood Cliffs, NJ: Prentice-Hall

van Buuren, Stef and Willem J. Heiser (1989) "Clustering N Objects into K Groups Under Optimal Scaling of Variables," Psychometrica, Vol. 54, n. 4, December, pp. 699 - 706

van der Lans, Ivo A. and Willem J. Heiser (1992) "Constrained part-worth estimation in conjoint analysis using the self-explicated utility model," International Journal of Research in Marketing, Vol. 9, n. 4, December, pp. 325 - 344

Vriens, Marco, M. Wedel, and T. Wilms (1992) "Simultaneous segmentation and estimation in conjoint models: A Monté Carlo comparison of recently proposed methods," Working paper, Faculty of Economics, University of Groningen

Wedel, Michel and Cor Kistemaker (1989) "Consumer Benefit Segmentation Using Clusterwise Linear Regression," International Journal of Research in Marketing, Vol. 6, n. 1, September, pp. 45 - 59

Wedel, Michel and Jan-Benedict E. M. Steenkamp (1989) "A fuzzy clusterwise regression approach to benefit segmentation," International Journal of Research in Marketing, Vol. 6, pp. 241 - 258

Wedel, Michel and Jan-Benedict E. M. Steenkamp (1991) "A Clusterwise Regression Method for Simultaneous Fuzzy Market Structuring and Benefit Segmentation," Journal of Marketing Research, Vol. XXVIII, n. 4, November, pp. 385 - 396

Wiley, James B. (1993) "A Strategy For A Priori Segmentation In Conjoint Analysis," in: McAlister, Leigh and Michael L. Rothschild (Eds.), Advances in Consumer Research, Vol. XX, Provo, UT: Association for Consumer Research, pp. 142 - 148

Wind, Yoram (1978) "Issues and Advances in Segmentation Research," Journal of Marketing Research, Vol. XV, n. 3, August, pp. 317 - 337

Wittink, D. R., L. Krishnamurthi, and D. J. Reibstein (1989) "The Effects of Differences in the Number of Attribute Levels on Conjoint Results," Marketing Letters, Vol. 1, n. 2, May, pp. 113 - 123

Wittink, Dick R. and Philippe Cattin (1989) "Commercial Use of Conjoint Analysis: An Update," Journal of Marketing, Vol. 53,July, pp. 91 - 96

Wittink, Dick R., David J. Reibstein, William Boulding, John E. Bateson, and John W. Walsh (1989) "Conjoint Reliability Measures," Marketing Science, Vol. 8, n. 4, Fall, pp. 371 - 374

Wittink, Dick R., Marco Vriens, and Wim Burhenne (1994) "Commercial Use of Conjoint Analysis in Europe: Results and Critical Reflections," International Journal of Research in Marketing, Vol. 11, n. 1, pp. 41 - 52

Wright, Peter and Mary Ann Kriewall (1980) "State-of-Mind Effects on the Accuracy with Which Utility Functions Predict Marketplace Choice," Journal of Marketing Research, Vol. XVII, n. 3, August, pp. 277 - 293

Wyner, Gordon A. (1992a) "Uses and Limitations of Conjoint Analysis — Part I," Marketing Research: A Magazine of Management and Applications, Vol. 4, n. 2, June, pp. 42 - 44

Wyner, Gordon A. (1992b) "Uses and Limitations of Conjoint Analysis — Part II," Marketing Research: A Magazine of Management and Applications, Vol. 4, n. 3, September, pp. 46 - 47

Yager, Ron and D. Filev (1994) "Generation of Fuzzy Rules by Mountain Clustering," Journal of Intelligent & Fuzzy Systems, Vol. 2, n. 3, September, pp. 209 - 219

Young, Shirley and Barbara Feigin (1975) "Using the Benefit Chain for Improved Strategy Formulation," Journal of Marketing, Vol. 39, n. 3, July, pp. 72 - 74

Zeithaml, Valarie A. (1987) "Defining and Relating Price, Perceived Quality, and Perceived Value," Report No. 87-101, Cambridge, MA: Marketing Science Institute, June

Zeithaml, Valarie A. (1988) "Consumer Perceptions of Price, Quality, and Value: A Means-End Model and Synthesis of Evidence," Journal of Marketing, Vol. 52, n. 3, July, pp. 2 - 22

APPENDIX I

A pretest was conducted to provide guidelines for the final design of the experiment in terms of

1. importance of attributes,
2. order effects,
3. conjectured interaction effects and (negative) correlations, and
4. familiarity as a possible covariate.

The questionnaire used for the pretest is provided at the end of this appendix on page 275.

1. Importance of Attributes

The pretest was conducted to elicit stated importance of ten candidate attributes (questions number two to eleven) with the intention to narrow down this list to about six to eight at two or three levels which is considered to be a good balance between demands for conjoint design and realism of respondent task before one may experience simplified decision strategies. The pretest also encouraged to state criteria a respondent would use but that were not included in the importance ratings (question number twelve). Table A1 on the following page provides the responses to the five-point category rating scales on the questionnaire. There were thirty (30) useful responses with respondent number fourteen (14) having two missing cells (attributes 'referent others' and price).

Responses are in the following order:

Referent others (A); Familiarity (B); Weight (C); Display Type (D); Screen Size (E);

Keyboard (F); Firm Reputation (G); Price (H); Battery Life (I); Additional Features (J); Performance (K); Order of Questions (O); No. denotes the respondent number.

Table A1: Means, Importance Ratings, and Order Effects (Raw Responses)

Morning Class

No.	A	B	C	D	E	F	G	H	I	J	K	O	
1	4	2	3	4	3	1	5	3	3	1	4	0	
4	2	2	3	4	3	4	4	5	4	5	5	0	
5	2	4	2	4	3	3	3	2	4	5	5	0	
8	2	5	4	4	3	3	3	5	5	5	5	0	
10	1	3	3	5	4	4	5	5	2	5	5	0	
11	2	2	2	2	2	3	2	4	3	3	3	0	
14		2	3	3	3	3	3		4	4	5	0	
15	3	2	4	4	5	4	3	4	3	5	5	0	
	2.29	2.75	3.00	3.75	3.25	3.13	3.50	4.00	3.50	4.13	4.63		Mean0a
2	2	3	5	4	4	3	4	5	5	5	5	1	
3	4	2	2	2	3	4	3	2	3	5	3	1	
6	3	2	3	3	3	2	4	5	4	2	4	1	
7	1	2	3	2	3	2	4	4	3	4	4	1	
9	2	2	4	4	4	4	4	3	4	4	4	1	
12	4	2	3	2	3	3	3	2	3	4	5	1	
13	1	2	4	2	4	3	3	5	5	3	4	1	
16	2	5	3	5	4	2	2	3	5	4	4	1	
17	5	2	5	4	5	4	5	4	3	5	4	1	
	2.67	2.44	3.56	3.11	3.67	3.00	3.56	3.67	3.89	4.00	4.11		Mean1a
Evening Class													
18	1	3	5	2	4	5	2	4	4	4	2	0	
21	1	4	2	4	4	5	3	4	3	2	5	0	
22	4	2	2	4	3	4	4	4	4	5	5	0	
25	1	2	3	3	3	2	3	4	3	4	4	0	
26	3	3	1	3	4	4	1	4	3	2	5	0	
29	1	2	5	4	4	5	4	5	4	4	5	0	
30	2	5	4	5	4	4	5	3	5	5	5	0	
	1.86	3.00	3.14	3.57	3.71	4.14	3.14	4.00	3.71	3.71	4.43		Mean0b

Order Coding:
0 = regular
1 = reverse code

No.	A	B	C	D	E	F	G	H	I	J	K	O	
19	3	5	2	2	4	5	5	2	5	5	5	1	
20	1	3	2	4	3	3	4	4	4	5	5	1	
23	3	3	4	4	4	4	5	4	4	4	5	1	
24	4	3	5	3	4	5	5	5	4	5	5	1	
27	2	2	4	2	3	3	3	3	4	2	2	1	
28	2	2	3	5	3	4	4	4	5	5	5	1	
	2.50	3.00	3.33	3.33	3.50	4.00	4.33	3.67	4.33	4.33	4.50		Mean1b
	2.07	2.88	3.07	3.66	3.48	3.63	3.32	4.00	3.61	3.92	4.53		Mean0
	2.58	2.72	3.44	3.22	3.58	3.50	3.94	3.67	4.11	4.17	4.31		Mean1
	2.33	2.80	**3.26**	3.44	3.53	3.57	**3.63**	3.83	3.86	4.04	4.42		Overall means

Means are adjusted for missing cells; in the following statistical analysis they are interpolated (price, referent other).

As is easily conveyed by Table A1, the least importance for a purchase decision about a laptop computer is attributed to referent others, i.e. to what others think about a specific laptop computer. This is somewhat surprising given the relative complexity of the product, the relatively high price, and the proliferation of product comparisons in trade journals which they consider an important service for their readers. Second, all other attributes are, on average, at least important (scale value three; see questionnaire). Therefore, none of those were dropped for the conjoint evaluation. However, 'referent others' was dropped as an attribute from further consideration.

The only other non-technical, user-referent attribute, firm reputation, scored a mid place in importance ratings. Thus it is included in the conjoint task and provides the manipulation for the user-referent attribute set (A2). Additional idiosyncratic decision criteria obtained with the last question on the pretest questionnaire resulted in no discernible broad categories in addition to the stated ones that may have been overlooked.

2. Order Effects

Order effects were tested using one-way ANOVA (and with total and six Y´s MANOVA) to see if special precautions are necessary for questionnaire layout and stimulus set construction. There are no significant order effects (Tukey-Kramer q*), neither with classes.

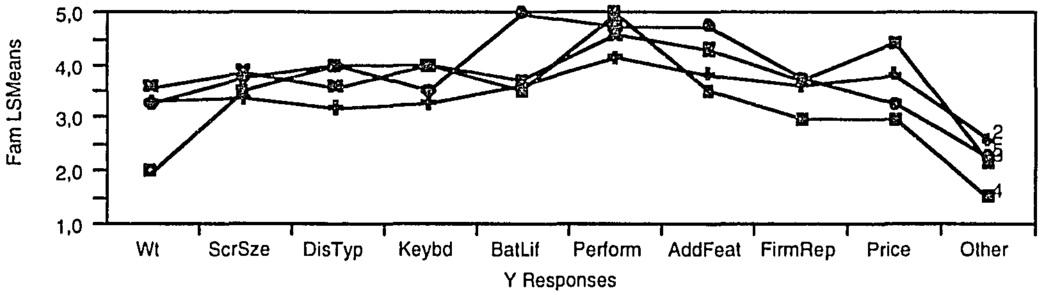
3. (Negative) Correlations

While there were several positive attribute correlations at the .5 level, only a slightly negative correlation between 'performance' and 'weight' was registered (-.141 with product moment and -.107 with rank correlation). However, partialled with respect to all other variables, this product moment correlation increased to a negative -.558. As Johnson, Meyer, and Ghose (1989) found adverse effects at a level of -.33 (p. 268), this interaction is tested.

4. Familiarity As Possible Covariate

Finally, a covariance analysis was conducted using familiarity with the product class in order to elicit this ancillary variable´s potential for revealing differentiating benefit attributions of respondents (i.e. act as a control variable for consumer differences). Though not significant, a visual inspection suggests a potential for those controls to serve as useful segmentation bases. See Figure A1 below.

Figure A1: Familiarity as Predictor for Importance



Fam	Wt	ScrSze	DisTyp	Keybd	BatLif	Perform	AddFeat	FirmRep	Price	Other
2	3,312	3,375	3,187	3,250	3,625	4,125	3,812	3,625	3,812	2,562
3	3,571	3,857	3,571	4,000	3,714	4,571	4,286	3,714	4,429	2,143
4	2,000	3,500	4,000	4,000	3,500	5,000	3,500	3,000	3,000	1,500
5	3,250	3,750	4,000	3,500	5,000	4,750	4,750	3,750	3,250	2,250

Pretest Questionnaire

Below are 12 easy questions concerning "laptop" or "notebook" computers. Please answer them on the scales provided below the questions.

1. How familiar do you consider yourself with laptop or notebook computers ? (Consider what you heard, read, or saw about them, or maybe used yourself.)

□	□	□	□	□
not familiar	somewhat familiar	quite familiar	occasional user	regular user

Imagine you considered buying a laptop or notebook computer.

Below are a list of general characteristics or product attributes that may be considered when choosing among different laptops or notebooks. Please, indicate how important these characteristics are for you by choosing one of the boxes that best describes the importance of the characteristic. Please, think for a few seconds before proceeding to the next item.

2. How important is the weight of the laptop or notebook ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

3. How important is the screen size of the laptop or notebook ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

4. How important is the display type (monochrome or color) of the laptop or notebook ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

5. How important is the keyboard size (regular or smaller) of the laptop or notebook ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

6. How important is the battery life of the laptop or notebook ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

7. How important is the performance or speed of the laptop or notebook ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

8. How important is the presence of additional features (as for instance connection ports, faxmodem, CD-ROM) ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

9. How important is the firm's reputation offering the laptop or notebook (well-known or national brand, no-name) ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

10. How important is the price of a laptop or notebook ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

11. How important is what others (your family, friends, colleagues, journals) think of a laptop or notebook ?

□	□	□	□	□
not important	slightly important	important	very important	essential characteristic

12. What other things (apart from those listed above) would you look at when considering the purchase of a laptop or notebook computer ? List whatever you would consider in purchasing a laptop or notebook computer.

Thank you very much for your cooperation. (RH)

APPENDIX II

Experimental design package.

(See stapled package at end.)

APPENDIX III

Sequence of Task Administration and Procedures

The questionnaires were administered in a classroom setting. Respondents took questionnaires home, and returned them within the next two class sessions. Instructors of six (6) classes in Winter Term 1995, and one (1) class in Spring Term 1995 gave their permission to administer the questionnaires within their classes.

First Replication

1) Introduction/Explanation

Right at the beginning, researcher, study purpose, and the type of information requested were introduced. Subjects were told that there are two administrations of the experiment. Then, those volunteering to participate were asked to raise their hands, and questionnaires were distributed. With subjects participating in the study proceedings were as follows (explanations were kept to the necessary minimum because of the danger of influence through explanation):

- Introduction of study purpose and required information package.
- Subjects are told that it is important they provide the information to the best of their knowledge ("Take your time.").
- Explanation and visualization of • (product) attributes, and
  • attribute levels

  (material is also in experimental package, viz. questionnaire).

- Framing with:

  "Imagine you are in the process of evaluating different laptop computers for potential purchase for yourself."
- Explanation of stimulus evaluation task.
- Explanation of scale
  - 7-point category-rating for controls;
  - 0 - 100 point likelihood of purchase scale for stimulus task.

2) Phase One/Self-Explicated

The first phases of tests request information concerning self-explicated ratings of the attribute levels, and the importance of attributes. This phase makes respondents familiar with the task and eases the evaluative phase of the conjoint task.

- Ask student name and class number.
- Ask control variable familiarity with task (category rating scale similar to pretest)
- Desirability rating of the levels per attribute on a 0 to 100 point rating scale.
- Quantitative judgment rating of importance of attributes on a 0 to 100 point importance scale with anchoring at the best attribute as suggested by Srinivasan (1988, p. 296).

3) Phase Two/ Conjoint Task

The conjoint calibration task consists of 27 stimuli, the ordering of which was randomized first, then this randomized order and a reverse order were used for the calibration. Warm-up profiles as suggested by Louviere (1988) were not provided, as subjects became familiar with the task when they provided the self-explicated ratings.

- 27 stimuli (randomized and reverse-ordered) are evaluated by rating their 'likelihood of purchase' on a 0 to 100 rating scale. Subjects were advised to first go through the profile descriptions and make themselves familiar with them, then rate these profiles on the scale.

4) Phase Three/ Control Variables / Ancillary Variables

Some additional demographic variables were collected, next.

- Gender (binary).
- Age.
- Student standing in years (undergraduate/graduate; freshman, sophomore, junior, senior).
- Work experience in years.
- Computer ownership in years.
- Computer usage and experience in years.

5) Phase Four/ Holdout Choice and Rating

The collection of holdout sample data consisted of a modified conjoint task in which ratings and choices were made for four sets of four stimuli each. This resulted in sixteen (16) evaluations as the holdout sample (4 x 4 choice sets). The pofiles for this task were constructed as a 2-level-extreme design as indicated in Appendix IV. The following data were collected per set:

- First choice hit; after evaluating the four profiles in a set, subjects were asked to choose the one that they would most likely purchase.
- Ratings of the stimuli of a set on the same 0 to 100 'likelihood of purchase' scale as before.

6) At end of task, two more control variable were asked and recorded:

- Record time to complete the experiment.
- Ask to rate (on a scale from 0 to 100) how difficult this task was.

Second Replication

Introduction; limited. Phase One, Phase Two, Phase Three, and Phase Four; as before (1 to 2 weeks after first replication). Debriefing.

APPENDIX IV/1

Fractional factorial coding structure for the two fractional factorials used for calibration in this study (FF1, FF1). Design $2^3 3^6$ according to Addelman 'Basic Plan 6' (1962, p. 38). For FF1 and FF2, there are three factors at two (2) levels and six factors at three (3) levels for a total of 27 stimuli. Levels of the factors are coded as follows (3-level ; 2-level):

Level One:	-1	;	-1
Level Two:	0	;	1
Level Three:	1		

Coding structure for FF1:

Stimulus 1:	-1	-1	-1	-1	-1	-1	-1	-1	-1
Stimulus 2:	-1	-1	-1	0	0	1	0	1	0
Stimulus 3:	-1	-1	-1	1	1	0	1	0	1
Stimulus 4:	-1	1	1	-1	-1	-1	0	0	0
Stimulus 5:	-1	1	1	0	0	1	1	-1	1
Stimulus 6:	-1	1	1	1	1	0	-1	1	-1
Stimulus 7:	1	-1	1	-1	-1	-1	1	1	1
Stimulus 8:	1	-1	1	0	0	1	-1	0	-1
Stimulus 9:	1	-1	1	1	1	0	0	-1	0
Stimulus 10:	-1	1	1	-1	0	0	-1	-1	0
Stimulus 11:	-1	1	1	0	1	-1	0	1	1
Stimulus 12:	-1	1	1	1	-1	1	1	0	-1
Stimulus 13:	1	1	-1	-1	0	0	0	0	1
Stimulus 14:	1	1	-1	0	1	-1	1	-1	-1
Stimulus 15:	1	1	-1	1	-1	1	-1	1	0
Stimulus 16:	1	1	-1	-1	0	0	1	1	-1
Stimulus 17:	1	1	-1	0	1	-1	-1	0	0
Stimulus 18:	1	1	-1	1	-1	1	0	-1	1
Stimulus 19:	1	-1	1	-1	1	1	-1	-1	1
Stimulus 20:	1	-1	1	0	-1	0	0	1	-1
Stimulus 21:	1	-1	1	1	0	-1	1	0	0
Stimulus 22:	1	-1	1	-1	1	1	0	0	-1
Stimulus 23:	1	-1	1	0	-1	0	1	-1	0
Stimulus 24:	1	-1	1	1	0	-1	-1	1	1
Stimulus 25:	-1	-1	-1	-1	1	1	1	1	0
Stimulus 26:	-1	-1	-1	0	-1	0	-1	0	1
Stimulus 27:	-1	-1	-1	1	0	-1	0	-1	-1

APPENDIX IV/2

Coding structure for FF2:

Stimulus 1:	1	1	1	-1	-1	-1	-1	-1	-1
Stimulus 2:	1	1	1	0	1	0	1	0	1
Stimulus 3:	1	1	1	1	0	1	0	1	0
Stimulus 4:	1	-1	-1	0	0	0	0	1	1
Stimulus 5:	1	-1	-1	1	-1	1	-1	-1	0
Stimulus 6:	1	-1	-1	-1	1	-1	1	0	-1
Stimulus 7:	-1	1	-1	1	1	1	1	0	0
Stimulus 8:	-1	1	-1	-1	0	-1	0	1	-1
Stimulus 9:	-1	1	-1	0	-1	0	-1	-1	1
Stimulus 10:	1	-1	-1	-1	-1	0	0	0	0
Stimulus 11:	1	-1	-1	0	1	1	-1	1	-1
Stimulus 12:	1	-1	-1	1	0	-1	1	-1	1
Stimulus 13:	-1	-1	1	0	0	1	1	-1	-1
Stimulus 14:	-1	-1	1	1	-1	-1	0	0	1
Stimulus 15:	-1	-1	1	-1	1	0	-1	1	0
Stimulus 16:	-1	-1	1	1	1	-1	-1	1	1
Stimulus 17:	-1	-1	1	-1	0	0	1	-1	0
Stimulus 18:	-1	-1	1	0	-1	1	0	0	-1
Stimulus 19:	-1	1	-1	-1	-1	1	1	1	1
Stimulus 20:	-1	1	-1	0	1	-1	0	-1	0
Stimulus 21:	-1	1	-1	1	0	0	-1	0	-1
Stimulus 22:	-1	1	-1	0	0	-1	-1	0	0
Stimulus 23:	-1	1	-1	1	-1	0	1	1	-1
Stimulus 24:	-1	1	-1	-1	1	1	0	-1	1
Stimulus 25:	1	1	1	1	1	0	0	-1	-1
Stimulus 26:	1	1	1	-1	0	1	-1	0	1
Stimulus 27:	1	1	1	0	-1	-1	1	1	0

APPENDIX IV/3

Fractional factorial coding structure for the holdout fractional factorial used in this study (FF-Holdout). Design 2^9 according to SAS-Institute (1994, Ch. 26).

For FF-Holdout, there are the three factors at two (2) levels and the six factors with three levels also at only two (2) levels for a total of 16 stimuli. Levels of the factors are coded as follows:

Level One:	-1
Level Two:	1

Coding structure for holdout profiles (16 treatments; 2-level extreme design; Resolution III: main effects):

Stimulus 1:	-1	-1	-1	-1	1	-1	-1	1	-1
Stimulus 2:	-1	-1	-1	1	-1	1	1	-1	1
Stimulus 3:	-1	-1	1	-1	-1	1	1	-1	-1
Stimulus 4:	-1	-1	1	1	1	-1	-1	1	1
Stimulus 5:	-1	1	-1	-1	-1	1	-1	1	1
Stimulus 6:	-1	1	-1	1	1	-1	1	-1	-1
Stimulus 7:	-1	1	1	-1	1	-1	1	-1	1
Stimulus 8:	-1	1	1	1	-1	1	-1	1	-1
Stimulus 9:	1	-1	-1	-1	-1	-1	1	1	1
Stimulus 10:	1	-1	-1	1	1	1	-1	-1	-1
Stimulus 11:	1	-1	1	-1	1	1	-1	-1	1
Stimulus 12:	1	-1	1	1	-1	-1	1	1	-1
Stimulus 13:	1	1	-1	-1	1	1	1	1	-1
Stimulus 14:	1	1	-1	1	-1	-1	-1	-1	1
Stimulus 15:	1	1	1	-1	-1	-1	-1	-1	-1
Stimulus 16:	1	1	1	1	1	1	1	1	1

APPENDIX IV/4

The first two fractional factorial coding structures used in this study were obtained from Addelman´s 'Basic Plan 6' (1962, p. 38) which provides more orthogonal codes than Connor and Young (1961, p. 40), and thus allows to use orthogonal polynomials without adjustments for estimation. Also, random sampling of a complete Resolution IV design is inferior to a constructed design where the desired interaction may be better controlled. 'Basic Plan 6' is appropriate as two different orthogonal fractional factorial designs are necessary, and as there are two-level and three-level factors mixed in the design. Admittedly, this mixed-level design complicates matter, but 'saves' three (3) degrees of freedom compared to a solely three-level design. Also, this design allows for estimation of the interaction between one two-level and one three-level factor, as required in the study (the one suggested important by the pretest).

The first two fractional factorials were obtained using the 4 x 3^6 design of the plan with those two correspondence schemes provided on p. 26. FF1 uses the first correspondence scheme and columns five (5) to ten (10) of 'Basic Plan 6'. FF2 uses the second correspondence scheme and columns eight (8) to thirteen (13) of 'Basic Plan 6'. Both plans´ profile orders were randomized, and the attribute 'Performance/Speed' was assigned to column A, while attribute 'Weight' was assigned to column D, allowing for estimation of this interaction. Then, attributes were ordered according to Table X on page 99. Levels were not assigned randomly to the profiles but ordered as assumed from least to most preferred. In the reverse-coded questionnaires, profiles and features are in reverse order. With two attribute sets (A1, A2), two fractional factorial plans (FF1, FF2), two orders (order, reverse-order), and timely procedure as in Figure 8 on page 92, six different questionnaires were devised and randomly assigned to respondents.

APPENDIX V

Sample of Survey Instruments

Laptop / Notebook Computer Study

Data solicited with this questionnaire is used for the sole purpose of research in consumer preference and choice behavior. Participation is completely voluntary. Your decision not to participate will not affect your grade.

All data collected henceforth in this questionnaire, will be kept confidential. This implies that only the researcher can identify each participant and his/her responses but assures that any data provided will be held private in a locked file cabinet and not be revealed to others. As soon as the second questionnaire has been matched with your name, any linkage of recorded data to a name or class number will be erased. All data is then only kept as anonymous data, and questionnaires are destroyed. From then on, there is no way to identify any respondent any more.

It will take about one (1) hour to answer the questionnaire. You may stop at any time into the questionnaire.

In the following questionnaire you are asked to provide information about "laptop" or "notebook" computers. You are asked to repeat this questionnaire in four (4) weeks in the same classroom setting.

The questionnaire is conducted in four (4) phases.

In Phase I, you are asked to

- provide your name and class number,
- indicate, how familiar you consider yourself with laptop or notebook computers,
- rate desirabilities of attribute levels, and
- rate the importance of attributes.

In Phase II, you are asked to rate twenty-seven (27) generic product profiles on a zero (0) to one hundred (100) 'likelihood-of-purchase' scale.

In Phase III, you are asked to provide the following information about yourself:

- Gender,
- Age,
- Undergraduate/Graduate and year in college,
- Years of work experience,
- Years of computer ownership,
- Years of computer usage and experience.

In Phase IV, you are asked to

- make choices, and
- rate another sixteen (16) product profiles on a zero (0) to one hundred (100) 'likelihood-of-purchase' scale.

Following are the questions concerning "laptop" or "notebook" computers. Please, answer them on the scales provided below the questions, or record your rating as indicated.

If you have concerns or questions about this study, please contact the Chair of the Human Subjects Review Committee, Office of Research and Sponsored Projects, 105 Neuberger Hall, Portland State University, 503/725-3417.

Phase I

_________________________ Name (First, Last)

__________ Class Number

1. How familiar do you consider yourself with laptop or notebook computers ? (Consider what you heard, read, or saw about them, or maybe used yourself.)

□	□	□	□	□
not familiar	somewhat familiar	quite familiar	occasional user	regular user

Imagine you considered buying a laptop or notebook computer, and you are in the process of evaluating different laptop computers for potential purchase for yourself.

Below are a list of general characteristics or product attributes, and their respective levels (i.e. possible ranges this attribute can assume), that may be considered when choosing among different laptops or notebooks.

Please, first examine the attributes and levels on the following page. Then, for each attribute, rate the desirabilities of the different levels within the attribute. Do not consider other attributes when rating desirabilities of levels within an attribute. Note the desirabilities of levels on a zero (0 = so undesirable a level that the whole product would be rejected) to one hundred scale (100 = attribute level is the most desirable; if considered alone, this would lead to sure buy).

Rate desirabilities of the attribute levels on a zero (0 = so undesirable a level that the whole product would be rejected) to one hundred scale (100 = attribute level is the most desirable; if considered alone, this would lead to sure buy). Please, take your time.

A. Weight

9 pounds ____________________

7 pounds ____________________

5 pounds ____________________

B. Screen Size

8.4 inch (diagonal) ____________________

9.4 inch (diagonal) ____________________

10.4 inch (diagonal) ____________________

C. Display Type

Monochrome ____________________

Color ____________________

D. Base Price

$ 3500 ____________________

$ 2500 ____________________

$ 1500 ____________________

E. Keyboard Size

Smaller than regular size ____________________

Regular size ____________________

F. Battery Life

3 hours ____________________

5 hours ____________________

7 hours ____________________

G. Performance/Speed

Comfortable for word-processing ____________________

Fast for big spreadsheet and imaging ____________________

H. Presence of Additional Features

No additional features ____________________

Expansion slots for keyboard, monitor, others ____________________

Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others ____________________

I. Pointing Device

Mouse ____________________

Trackball ____________________

Trackpad ____________________

For each of the attributes, consider: If there were one attribute for which it would be most important for you to get the best level, which attribute would that be ? Assign 100 points to this 'critical attribute' (if there is more than one 'critical attribute' assign all of those 100 points). Now consider each of the remaining attributes. For each attribute, how important is it for you to get the best level of this attribute ? If it is only half as important for this attribute as for the 'critical attribute'(s), assign it 50 points ... In general, assign zero (0) to one hundred (100) points to reflect how important it is on this attribute (compared to the 'critical attribute') to have the best level instead of the worst. Please, take your time before proceeding to the next item.

A.	Weight	____________	9 pounds 7 pounds 5 pounds
B.	Screen Size	____________	8.4 inch (diagonal) 9.4 inch (diagonal) 10.4 inch (diagonal)
C.	Display Type	____________	Monochrome Color
D.	Base Price	____________	$ 3500 $ 2500 $ 1500
E.	Keyboard Size	____________	Smaller than regular size Regular size
F.	Battery Life	____________	3 hours 5 hours 7 hours
G.	Performance/Speed	____________	Comfortable for word-processing Fast for big spreadsheet and imaging
H.	Presence of Additional Features	____________	No additional features Expansion slots for keyboard, monitor, others Faxmodem, CD-ROM, expansion slots for keyboard, monitor, others
I.	Pointing Device	____________	Mouse Trackball Trackpad

Phase II

In this section of the questionnaire you are asked to rate twenty-seven (27) generic product profiles on a zero (0) to one hundred (100) 'likelihood-of-purchase' scale. You may first take a look at the product profile before rating the first one. It is important that you take your time for each profile description (about one (1) minute, each). Please, rate a profile on a 'likelihood-of-purchase' scale reaching from zero (0 = under any circumstances definitely would not buy) to one hundred (100 = certainly would buy). Then, proceed to the next item page.

Profile # 1

Weight:	**7 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 1500**
Keyboard Size:	**Regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 2

Weight:	**5 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 3500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackball**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 3

Weight:	**9 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 1500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 4

Weight:	**9 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 3500**
Keyboard Size:	**Regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Trackball**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 5

Weight:	**9 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 2500**
Keyboard Size:	**Regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 6

Weight:	**7 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 2500**
Keyboard Size:	**Regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Trackball**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 7

Weight:	**7 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 2500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 8

Weight:	**5 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 1500**
Keyboard Size:	**Regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Trackball**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 9

Weight:	**5 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 2500**
Keyboard Size:	**Regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 10

Weight:	**5 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 3500**
Keyboard Size:	**Regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 11

Weight:	**9 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 3500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 12

Weight:	**9 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 1500**
Keyboard Size:	**Regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 13

Weight:	**9 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 2500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 14

Weight:	**5 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 2500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 15

Weight:	**7 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 1500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackball**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 16

Weight:	**5 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 2500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackball**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 17

Weight:	**5 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 1500**
Keyboard Size:	**Regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 18

Weight:	**5 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 3500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 19

Weight:	**7 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 3500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackball**

zero (0) = under any circumstances definitely would not buy	[Note number] **Likelihood of Purchase**	hundred (100) certainly would buy

Profile # 20

Weight:	**9 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 2500**
Keyboard Size:	**Regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackball**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 21

Weight:	**7 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 3500**
Keyboard Size:	**Regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 22

Weight:	**7 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 2500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 23

Weight:	**7 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 3500**
Keyboard Size:	**Regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 24

Weight:	**9 pounds**
Screen Size:	**10.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 1500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**7 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**
Pointing Device:	**Trackball**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 25

Weight:	**9 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Monochrome**
Base Price:	**$ 3500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Comfortable for word processing**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 26

Weight:	**5 pounds**
Screen Size:	**9.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 1500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**3 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**No additional features**
Pointing Device:	**Trackpad**

zero (0)	[Note number]	hundred (100)
= under any circumstances definitely would not buy	**Likelihood of Purchase**	certainly would buy

Profile # 27

Weight:	**7 pounds**
Screen Size:	**8.4 inch (diagonal)**
Display Type:	**Color**
Base Price:	**$ 1500**
Keyboard Size:	**Smaller than regular size**
Battery Life:	**5 hours**
Performance/Speed:	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**Expansion slots for keyboard, monitor, others**
Pointing Device:	**Mouse**

zero (0)	[Note number]	hundred (100)
= under any		certainly
circumstances	**Likelihood**	would
definitely would	**of**	buy
not buy	**Purchase**	

Phase III

For the following questions, please mark the box that applies to you.

2. Gender: □ Female □ Male

3. Age: ________ Years

4. Student Status: □ Graduate □ Undergraduate

5. Years in College: ________ Years

6. Years of Work Experience: ________ Years

7. Years of Computer Ownership: ________ Years

8. Years of Computer Usage and Experience: ________ Years

Phase IV

On each of the following four (4) pages you will find four (4) sets of product profiles listed side by side. For each set of four (4) product profiles, examine the four profiles carefully and do the following:

1. Choose the best out of four (4) product profiles by marking the box below your choice.
2. Rate each of the four (4) profiles on a page on a 'likelihood-of-purchase' scale reaching from zero (0 = under any circumstances definitely would not buy) to one hundred (100 = certainly would buy).

	Profile # 1	Profile # 2	Profile # 3	Profile # 4
Weight:	**9 pounds**	**9 pounds**	**9 pounds**	**5 pounds**
Screen Size:	**10.4 inch (diagonal)**	**8.4 inch (diagonal)**	**10.4 inch (diagonal)**	**10.4 inch (diagonal)**
Display Type:	**Color**	**Monochrome**	**Color**	**Monochrome**
Base Price:	**$ 3500**	**$ 1500**	**$ 3500**	**$ 3500**
Keyboard Size:	**Smaller than regular size**	**Smaller than regular size**	**Regular size**	**Smaller than regular size**
Battery Life:	**7 hours**	**3 hours**	**3 hours**	**7 hours**
Performance Speed:	**Fast for big spreadsheet and imaging**	**Fast for big spreadsheet and imaging**	**Comfortable for word processing**	**Fast for big spreadsheet and imaging**
Presence of Additional Features:	**No additional features**	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**	**Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others**	**No additional features**
Pointing Device:	**Trackpad**	**Trackpad**	**Trackpad**	**Mouse**
Best: (Mark One)	□	□	□	□
Rate: **(Likelihood)** **(of)** **(Purchase)**	___	___	___	___

(0 = under any circumstances definitely would not buy — 100 = certainly would buy)

	Profile # 5	Profile # 6	Profile # 7	Profile # 8
Weight:	9 pounds	5 pounds	9 pounds	5 pounds
Screen Size:	10.4 inch (diagonal)	8.4 inch (diagonal)	8.4 inch (diagonal)	8.4 inch (diagonal)
Display Type:	Monochrome	Color	Color	Color
Base Price:	$ 1500	$ 1500	$ 3500	$ 1500
Keyboard Size:	Regular size	Regular size	Regular size	Smaller than regular size
Battery Life:	7 hours	7 hours	3 hours	3 hours
Performance Speed:	Fast for big spreadsheet and imaging	Comfortable for word processing	Fast for big spreadsheet and imaging	Fast for big spreadsheet and imaging
Presence of Additional Features:	Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others	No additional features	No additional features	Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others
Pointing Device:	Mouse	Mouse	Mouse	Mouse
Best: (Mark One)	□	□	□	□
Rate: (Likelihood) (of) (Purchase)	——	——	——	——

(0 = under any circumstances definitely would not buy — 100 = certainly would buy)

	Profile # 9	Profile # 10	Profile # 11	Profile # 12
Weight:	5 pounds	9 pounds	5 pounds	5 pounds
Screen Size:	10.4 inch (diagonal)	8.4 inch (diagonal)	8.4 inch (diagonal)	10.4 inch (diagonal)
Display Type:	Color	Color	Monochrome	Monochrome
Base Price:	$ 1500	$ 3500	$ 3500	$ 3500
Keyboard Size:	Regular size	Smaller than regular size	Smaller than regular size	Regular size
Battery Life:	7 hours	7 hours	7 hours	3 hours
Performance Speed:	Fast for big spreadsheet and imaging	Comfortable for word processing	Comfortable for word processing	Comfortable for word processing
Presence of Additional Features:	Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others	Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others	Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others	Faxmodem, CD-ROM, Expansion slots for keyboard, monitor, others
Pointing Device:	Trackpad	Mouse	Trackpad	Mouse
Best: (Mark One)	□	□	□	□
Rate: (Likelihood) (of) (Purchase)	____	____	____	____

(0 = under any circumstances definitely would not buy — 100 = certainly would buy)

	Profile # 13	Profile # 14	Profile # 15	Profile # 16
Weight:	5 pounds	9 pounds	5 pounds	9 pounds
Screen Size:	10.4 inch (diagonal)	8.4 inch (diagonal)	8.4 inch (diagonal)	10.4 inch (diagonal)
Display Type:	Color	Monochrome	Monochrome	Monochrome
Base Price:	$ 1500	$ 1500	$ 3500	$ 1500
Keyboard Size:	Smaller than regular size	Regular size	Regular size	Smaller than regular size
Battery Life:	3 hours	7 hours	3 hours	3 hours
Performance Speed:	Comfortable for word processing	Comfortable for word processing	Fast for big spreadsheet and imaging	Comfortable for word processing
Presence of Additional Features:	No additional features	No additional features	No additional features	No additional features
Pointing Device:	Trackpad	Trackpad	Trackpad	Mouse
Best: (Mark One)	□	□	□	□
Rate: (Likelihood) (of) (Purchase)	——	——	——	——

(0 = under any circumstances definitely would not buy — 100 = certainly would buy)

9, Please, record the time you returned the questionnaire.

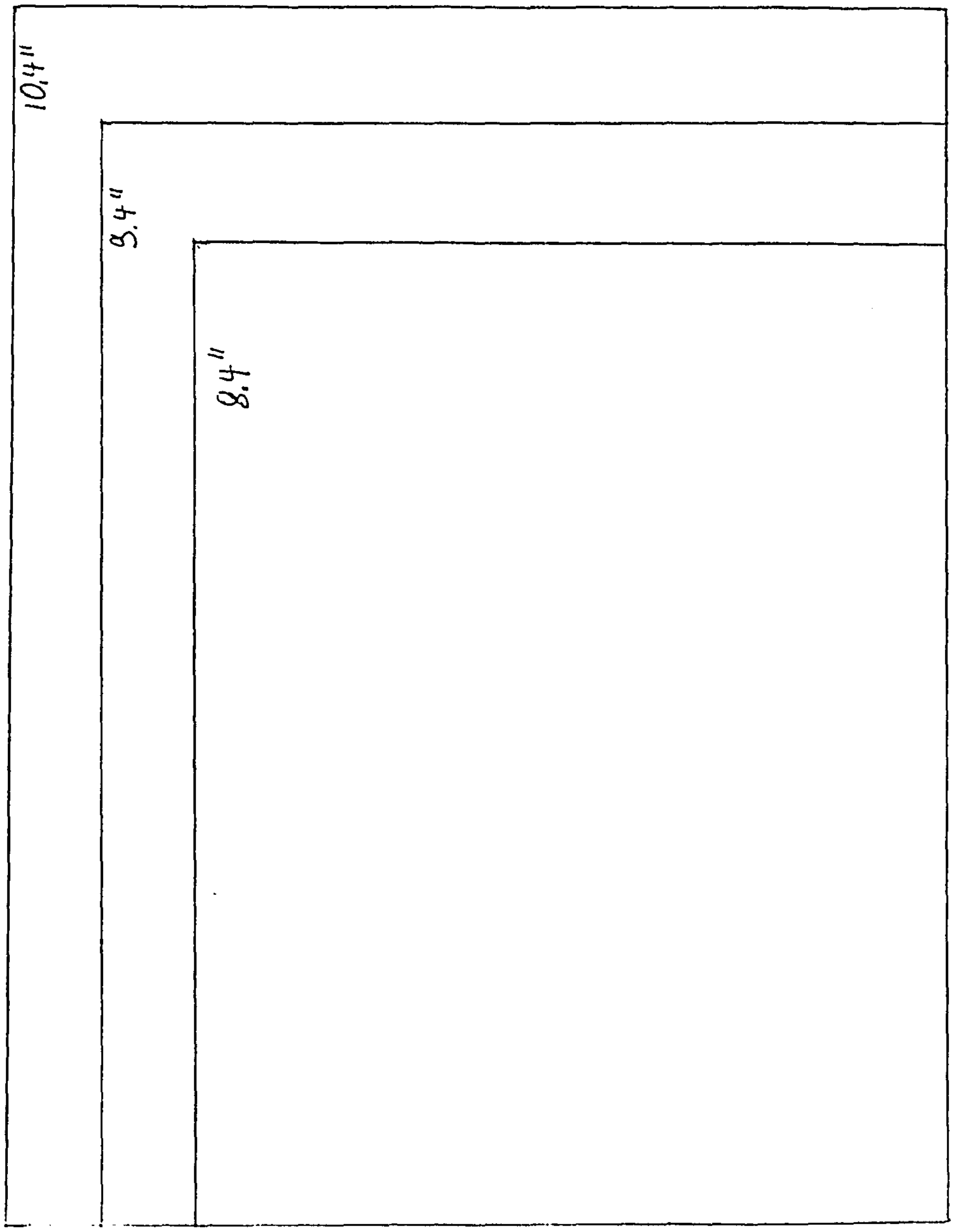
____
Time
(hour : minute)

10. Please indicate how difficult a task the above series of questions has been to you on the scale provided below (mark one).

□	□	□	□	□	□	□
not difficult at all	slightly difficult	moderately difficult	difficult	very difficult	extremely difficult	impossible task

11. Please, feel free to put any comments or remarks concerning the questionnaire, the product, the administrator, administration of the task, or yourself on the space provided below.

Thank you very much for your cooperation. (RH)


10.4"
9.4"
8.4"