RJ10397 (A0611-046) November 14, 2006 (Declassified 11/13/2007)
Other

IBM Research Report

Causal Models for Business Decision Making: An Empirical Investigation into the Critical Success Factors that Drive Sales

Sunil J. Noronha, Joseph D. Kramer
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Causal Models for Business Decision Making:

An Empirical Investigation into the Critical Success Factors that Drive Sales

Sunil J. Noronha and Joseph D. Kramer

IBM Almaden Research Center

San Jose, California

Abstract

Managerial decision makers need reliable methods for achieving business objectives, whether they seek to maximize revenue, profit, or some other measure of business performance. However, businesses as well as their external environments are quite complex and difficult to analyze, and in practice it is hard to predict the results of a given managerial action. For example, if a retailer spends a million dollars redesigning the floor layout of their stores, will the retailer at least break even in terms of increased sales? Or would it be wiser to invest in improving the quality of service from the salespeople? For guidance in making these investment decisions, the business community currently relies at best on unreliable correlational models, or on gut instinct. There has been a lack of sufficiently powerful scientific methods for building models that can reliably predict the results of a business strategy, and thus enable the decision maker to maximize return on investment.

However recent developments in the emerging field of causal modeling offer substantial advances that can change the practice of managerial decision making. Unlike the traditional `predictive' models commonly used in marketing, causal modeling techniques are used to reliably discover, represent, and reason about cause-and-effect relationships instead of correlations. The resulting quantitative models support estimation of the true expected returns from a given business intervention, enable computation of the optimal set of interventions, and support what-if analyses to handle uncertain situations. Exploiting the full power of the new approach requires integration of techniques from many disciplines, not just statistics. This paper presents such a unified framework for modeling decision processes and making the best investment decisions.

Our general framework was developed and tested in the process of building a specific causal model of retail sales, which we also present in this paper. The model captured numerous factors that affect a consumer's choice of retailer from whom they purchased a large electronics product. Not only did the model answer the above question about the effect of improving store layout---and many other questions relating to product, price, promotion, etc.---but it also showed how the retailer could double their sales via just three factors among the thousands studied.

The framework presented in the paper is used as the basis for introducing several state-of-the-art methods and tools for gaining customer insight. It shows how to rigorously use ethnographic techniques in support of variable discovery and causal modeling. It presents methods for fusing qualitative and quantitative research methods with scientific rigor, avoiding the weaknesses of each. It describes new statistical algorithms and tools for causal discovery. The paper also makes significant theoretical contributions by introducing the notion of quasi-deterministic relationships which are a fundamental type of choice behavior.

Keywords: Customer insight, decision making, business strategy, business investments, sales, consumer purchasing, retail, critical success factors, experimental study, ethnography, qualitative analysis, structural equation modeling, causal modeling, advertising and branding, pricing, store experience

Table of Contents

1 Introduction

> *What are the critical success factors that drive the sales of a product or service? What investment strategy should a business pursue to maximize profit?*

These are some of the central questions in Marketing and Business research. In response, many theories are routinely created and advocated by business researchers and consultants, often organized around a popular theme-of-the-day. For example it is widely believed that `customer-centricity' and `multi-channel retailing' should be the basis for business strategy, and that correctly designing customer `experiences' and `touch points' is key to influencing customer perception and thereby closing a sale (e.g., see [Lohse 2000, Rajgopal et al. 2001]). In retail store environments, `store format' and merchandizing are considered to be extremely important for driving sales, whereas in online environments, the `Web experience' is believed to be critical. Similarly, in other industries, e.g., in IT outsourcing, the client-provider relationship may be hypothesized to be the critical factor for successfully closing a deal.

However, while there are numerous theories about the critical success factors for any business outcome, these theories usually lack robust empirical support, a fact that is occasionally admitted in the literature. For example, consider *customer satisfaction* which is an extremely popular metric of business success. Although it may be *correlated* with sales, does it actually *cause* sales to go up? Does it induce `loyalty' and drive subsequent purchases? Gupta and Zeithaml [2005] found scant evidence for any causal link between customer satisfaction and retention or firm performance. Even when the factors under consideration are believed to truly influence the outcome, the true magnitudes of the effects are usually unknown. For example, is *customer experience* as important as *price*, or can slightly lower prices compensate for a poor experience? Will sales go up a lot, or only a little? Even if changing the design of a store or a Web site improves sales, since the change is an expensive proposition costing millions of dollars, are the gains worth the expense? These are important decisions that have major impact on business performance. Yet, for lack of dependable answers, senior executives normally make these decisions relying on their intuition rather than on deep scientific insight into the workings of their business.

One of the most difficult issues in this space of problems is pinning down cause and effect between the variables. Suppose we observed that *advertising a product in a flyer* is correlated with increased *sales* of the product, how do we know that there isn't an alternative explanation that explains away the apparent link, e.g., the same products are also placed *on sale in the store*, and customers tend to buy more of the product because it is on sale, not because they saw flyers? Almost all marketing studies consider one subgroup of variables in isolation (e.g., advertising variables, or store characteristics), as Gupta and Zeithaml [2005] point out. In the absence of experimental controls on these variables---and in practical business situations it is usually impossible to exert experimental control---or statistical adjustments for other potentially confounding variables, these studies leave unresolved the question of whether an observed link is a genuine causal link. If a link turns out not to be a causal one (or if the true causal effect is much smaller than that implied by a large correlation), a business investment in the hypothesized control variable becomes a bad decision since it will not produce the expected returns in terms of sales. Thus the general uncertainty about the role and relative importance of each factor creates difficulties in designing a business strategy---it promotes speculation and guesswork instead of guiding strategy with a definitive and trustworthy model of how the business, its customers, and its environment work.

Part of the difficulty in pinning down the truly critical factors is that causality has traditionally been viewed as a philosophical concept. Although on one hand young children seem to understand cause and effect ("dropping the glass causes it to break") and physics relies on it routinely ("applying a force causes acceleration, not the other way around"), there has been no scientific formalization of the concept of causality itself. In contrast with the above deterministic examples, if we consider more complex and uncertain examples such as understanding the economic effects of a tax policy, or the psychological effect of a price incentive on purchase behavior, the standard warning comes to mind: "correlation is not causation", but there is no further assistance in how we should draw causal inferences. (So what do we *do* about that tax policy?). In the absence of formal tools, researchers have routinely relied on their intuition for guidance in drawing conclusions. The adverse results are evident every time we read a paper that presents a path-analytic model with variable *A* pointing to variable *B*, and we notice that reversing the link makes equal sense, but the authors don't provide a good defense for their choice of directionality. This usually leaves the discerning reader with the disconcerting feeling that most of these studies present the authors' favored views, but alternative theories are equally plausible, and trustworthy conclusions cannot be drawn. Some researchers have gone so far as to call the resulting contradictions "junk science" [Goertzel 2002].

We believe that this situation is now poised for a major change. Over the last decade, seminal work by Pearl [2000], Spirtes et al. [2000], and others has resulted in the placing of causality on a firm mathematical foundation. Part of the insight comes from the recognition that the traditional mathematical tools, e.g., algebra and probability theory are inherently symmetric (e.g., $F = ma$ can be written equivalently as $a = F/m$), and we need additional tools to capture and utilize the asymmetric information need to reason about causality (F causes a; a does not cause F). Combining statistics with the theory of directed graphs and a calculus of interventions provides a big step in this direction. The resulting progress is not just mathematical; there is also a major advance in modeling, because the conceptualization of causality connects really simple and intuitive concepts (e.g., blocking, intervention) with the less intuitive (e.g., conditional independence, exogeneity) but very powerful apparatus of statistics. Any linguistic advance that enables non-mathematicians to utilize powerful mathematical tools can have a profound effect. These advances also enable automated discovery of causal relationships from quantitative data. All of this can raise the art and science of building explanatory models to a whole new level, and necessitates the development of new research methodologies and analytic frameworks that exploit these advances.

We will introduce such a framework in the next section. A result of several years of real-world experimentation and development, we used the framework to build a cause-and-effect model of the factors driving sales for a large retail client. We will present this study to illustrate and ground our approach. The client was interested in understanding the rapidly changing Home Electronics market, and in particular, TV purchasing. The current societal transition from an analog to a digital world, the constant waves of new digital products, and the confluence of technologies like computers and TVs are rapidly changing the entire marketplace. There was little to no understanding of the role of these products in the consumer's world, how consumers view and feel about these products, what kinds of events triggered their interest in purchasing such products, what kinds of shopping processes and decision processes consumers went through to select a product or to select a retailer from whom to buy the product, and what influenced their choices. The retailer wanted to gain the kind of deep insight about consumers and the electronics business that would enable them to make the right merchandising decisions, price products correctly, communicate in the right manner with potential shoppers, and ultimately to influence the consumer's purchasing process in a way that maximally increased sales.

The study was exceptional in its coverage of factors that could potentially influence purchasing decisions. In particular, it covered customer thoughts and feelings (the space of `attitudes' and `values'), their environment and context, and all relevant areas of marketing, e.g., product attributes, pricing and promotions, store location, store layout, salespeople, advertising, perceptions of retailers; indeed the Quantitative Phase component of the study acquired a dataset containing over 1500 variables and over a million data points from shoppers across the United States. The study was also exceptional in terms of the detailed Discovery Phase research that preceded the quantitative study; a sample of shoppers who really were in the market to buy a new TV, and at a very early stage of awareness were recruited and their shopping experiences tracked by a team of ethnographers for over 8 months. This produced an extremely rich and detailed base of qualitative data on which the subsequent quantitative model was based. Because of its scale, the study provides the first comprehensive look in the marketing literature at how the multitude of factors play out against each other, which factors in the end affect sales significantly, and which factors only play a cameo role without a major effect on business outcomes.

One of the numerous insights derived from the study was that changes in Store Layout turned out to have little or no effect on sales. It is interesting to note that store design is one of the central preoccupations of retailers today. Millions of dollars are regularly spent by retailers to develop and deploy new store formats, in the hope that a better design will make consumers more likely to buy from the store. Although we did find a correlation between store layout variables and sales, it turned out that there was an alternative explanation as will be explained later in this report, and the correlation was therefore spurious. Brand turned out to behave in a similar fashion---although our client felt certain, as are many other marketers today, that stocking the right brands of product is one of the top five critical factors for sales. Our study showed that the correlation of brand with sales, which was indeed present in the data, was actually spurious and was explained away by other factors. The point to note here is that the exceptional coverage of our study was critical to finding the other variables that disentangled cause from

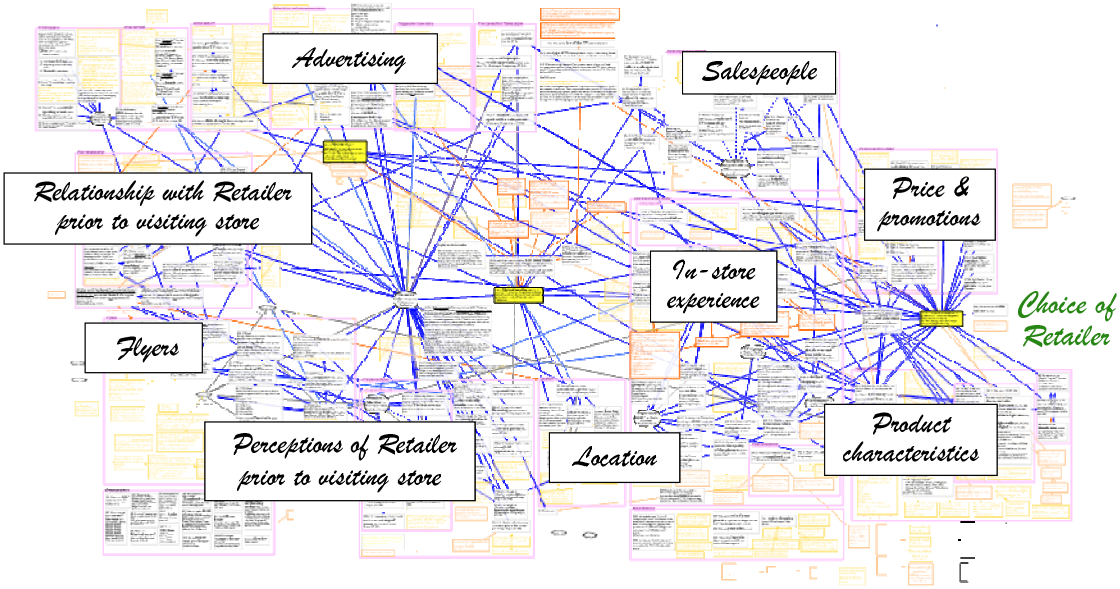


Figure 1 Snapshot of a causal model: retail sales of high-end televisions.

The superimposed annotations provide a rough description of the subject matter covered by the hundreds of variables depicted. The primary outcome variable on the right is the customer's choice of retailer from whom they will buy their television. The other major hub in the center is a variable that turned out be an extremely important intermediate outcome: the customer's decision to visit (or not visit) a given retailer. The variables are laid out so that cause and effect flows roughly from left to right.

correlation, thus overcoming the traditional weakness of most marketing studies. A narrower study would have missed the confounding variables.

It is also important to note that achieving such thorough coverage required substantial advances in research methodology, notably in the disciplines of ethnography, psychometrics, and causal modeling (the last being a new mathematical discipline [Pearl 2000] that is a descendant of traditional areas of statistics such as path analysis and structural equation modeling), and also in the rigorous integration of these disciplines which have been traditionally unconnected. These methodological insights in turn led us to rethink our entire approach to designing business strategies, ultimately developing a clean and intuitive framework with a solid foundation in the science of causal modeling. Much of this paper is devoted to explicating this framework and the underlying methods (Sections 2, 3.1 and 4.1), with the expectation that business managers who a desire a highly trustworthy basis for making investment decisions, or particular clarity of understanding about the functioning of their business and their customers, will capitalize on this approach.

The paper also presents several important findings discovered during the process of building our causal model of consumer purchasing. Scholars of decision theory and discrete choice modeling will be interested in the "quasi-deterministic patterns", a partly-deterministic partly-stochastic causal relationship between variables, which can be used to faithfully capture many aspects of human choice behavior. These choice models are introduced in Section 3.2.3, and their mathematical treatment is developed in Section 4.1.4.

Researchers familiar with causal modeling techniques will find Section 4.1.2 (and indeed most of Section 4) of particular interest, since we introduce new and improved algorithms for building causal models. Despite the seminal theoretical work of Pearl, Spirtes and others, there has been limited practical success in applying the theory to real-world problems, and indeed there is considerable controversy about whether causal discovery techniques work at all [Freedman 1993, 1998]. The controversy is not just about the validity of the theoretical assumptions behind the techniques, but also about the sanity of the models generated using these techniques; indeed we found that the only widely available software package for causal modeling, Tetrad [Spirtes et al. 2004], produced nonsensical models when tested on our data. Our investigation uncovered several flaws in the techniques used by the existing algorithms, and we were able to correct these problems via a combination of theoretical improvements and practical heuristics. The resulting toolkit that we developed generated excellent causal models. Therefore both theoreticians and practitioners who want to make these techniques perform successfully on real data will find the details of our methodology interesting and useful. Other quantitative researchers, especially in data mining, will also be interested in our critique of traditional methods for predictive modeling in Section 4.1.1.

Although the focus of this paper is on explicating a general framework for decision making via causal models, we do highlight a few findings specific to our retail application because of their importance to marketing researchers. For example, we repudiate common wisdom about the importance of *brand* and *store layout* (Section 4.4.1), and we describe the complex nonlinear models required to capture the influence of *product attributes* (Section 4.4.3.1). These and other highlights are presented in Sections 3.2 and 4.4.

The overall structure of this paper follows the chronological flow of our methodology and empirical study. The following section provides an overview of our framework, and then Sections 3 and 4 describe the two major phases of our study, while explaining the framework in greater detail. Finally, Section 5 summarizes the contributions of the study and suggests avenues for future research.

2 Overall Framework and Research Methodology

Building a model that explains purchasing behavior and gives a firm greater control over sales is challenging for a number of reasons. It is quickly evident that numerous factors must be modeled or at least examined for their potential to play a prominent explanatory role in the model; we have loosely grouped them into three classes in Figure 2.

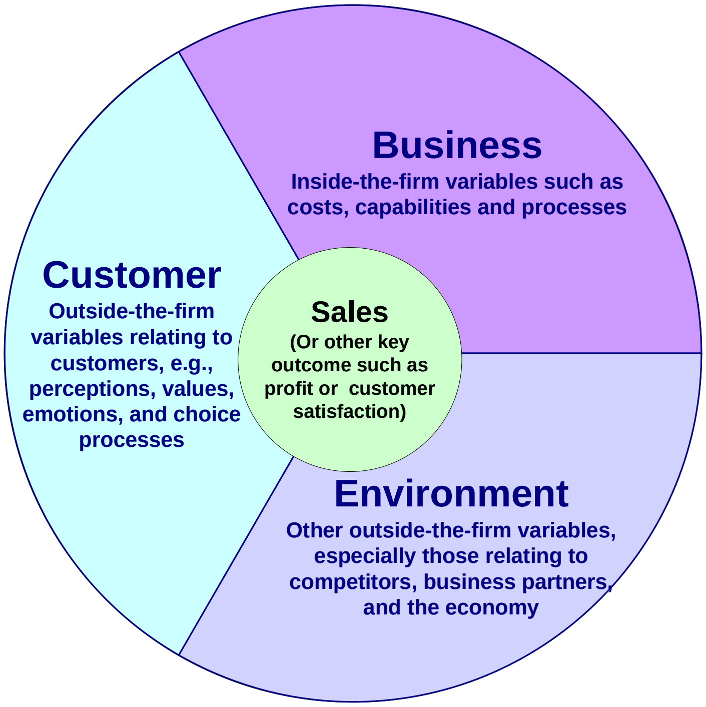

Business
Inside-the-firm variables such as costs, capabilities and processes
Customer
Outside-the-firm variables relating to customers, e.g., perceptions, values, emotions, and choice processes
Sales
(Or other key outcome such as profit or customer satisfaction)
Environment
Other outside-the-firm variables, especially those relating to competitors, business partners, and the economy

Figure 2. Three groups of variables that potentially have a causal effect on sales

The following are examples of variables from the three classes: A price-sensitive customer may be more likely to choose to visit a discounter such as Wal-Mart rather than a high-end electronics retailer such as Tweeter. The pricing strategy used by the business may be an everyday-low-price, or placing selected TVs on sale every week. And the competitors may offer price matching on identical TVs, or they may choose to stock very similar-looking but stripped down TV models at lower prices. Each of these variables can potentially change the ultimate outcome of where the customer chooses to buy their TV, and thus affects the sales of each retailer.

In a sense, the customer category is central to the model since each customer's decision is key to determining the outcome (a retailer's sales or share of market). However, it is clear that the other groups of variables influence the customer at numerous points along the decision process, and therefore must appear as causal variables in the model, at least when they are determined to have more than a trivial effect. This poses the methodological challenge: how do we acquire such wide-ranging data simultaneously covering all three categories of variables?[1] E.g., how do we know what the customer is

[1] If the data on all the variables to be modeled is not captured simultaneously it becomes extremely difficult to establish causal relationships between the variables. For example, if we obtain product availability information from retailers at one point in time, and store-visit information from consumers at another point in time, we have no idea what products were seen by the consumers who visited the store, and therefore cannot draw inferences about the effects of product attributes on purchase.

It is not strictly necessary to obtain end-to-end flows of cause and effect (e.g., to measure variables starting with a shopper's earliest awareness of need and finishing with the purchase) because of the local nature of causality (Section 4.1.2.3). If we get snapshots describing how a bunch of variables affect a proximal thinking or feeling, what are the specific products/price/merchandising decisions created by the business that affect the customer, and what are competitors doing to change the outcome? Not only do we need to first determine which of the infinitely many *potential* variables are actually worthy of study, we need to acquire data on these variables, from diverse sources, in a way that facilitates causal analysis.

These methodological challenges are further compounded by a couple of factors. While pricing and other `hard' business information can potentially be obtained from our retail client, `soft' customer information such as thoughts and feelings is hard to acquire. For that matter, it is not even clear how to conceptualize such variables and what exactly we should measure. Secondly, while information about our client's business may be relatively easily obtained directly from them, information about their competitors is hard to access for obvious reasons. Indeed these two challenges are not just methodological challenges for our research study; they constitute two traditional blind spots for our client: the inability to gain insight into their customers, and to understand how the business is performing against competitors (apart from the obvious end-results such as differences in market share). The prevalence of these blind spots explains the preoccupation with internal business metrics and processes that we have observed in the retail industry; this is the one area into which they can see. It follows that throwing light on outside-the-firm factors, especially the customer factors, would provide the greatest value to our clients, and therefore achieving great depth in this area has been a primary focus of our methodology, although in theory the methodology applies equally well across all three categories of variables.

This goal of gaining an understanding of customer thoughts and feelings, in a rigorous unbiased manner suitable for building scientific models of purchasing behavior, poses many more methodological challenges, some of which are listed in Figure 3. First it is hard to determine what to model; what kinds of data should we capture? The marketing literature is rife with theories and constructs, but a close examination of quickly reveals that the empirical underpinnings of those constructs are usually weak and their *causal* links to each other are either weak or unknown. For example, do people really carry `attitudes' and `values' in their heads, or does the brain process some other constructs? How are `attitudes' different from `beliefs' or `facts'? How are `consumer values' different from `emotions'? Do we need to model both? How do these interact with each other and how do they affect behavior? *Do* they actually affect behavior? E.g., does any human being ever `evaluate' their `satisfaction' with a product *before* they purchase it (as opposed to doing so when prompted to do so by a post-purchase survey), and does this presumed `evaluation' causally affect their purchase behavior? Or is a forced retrospective evaluation merely correlated with purchase behavior, and the behavior itself is driven by other causes, possibly including `emotions'? It is particularly surprising that a concept as central and commonly-used as `customer satisfaction' has been identified to

outcome, we can piece those snapshots together into a full model, even if the snapshots come from different shoppers. However his can be tricky, since we need some basis to assume that a downstream piece of the model is causally independent of the upstream pieces conditional on the intermediate variables. It is usually hard to justify the implicit conditional independence relationships, and only in rare circumstances can we study groups of variables in isolation and build models piecemeal. Given the wide diversity of the types of variables mentioned in Figure 2, simultaneous data collection is quite a challenging problem.

Figure 3 Key questions that must be answered by a methodology for business decision making based on causal models

1. How do we identify all the `potential causes’---the variables that might turn out to be important for affecting the business outcome? How do we minimize the risk of omitting critical factors?
2. Since we usually have to study people---customers and other stakeholders---how do we get inside their heads and find out what they are thinking and feeling?
3. What are the effects of those thoughts, feelings, activities, and contextual events on behaviors and subsequent outcomes? What are the mechanisms at play?
4. How do we scale up and obtain large quantities of such data, suitable for statistical analysis? How do we collect it reliably?
5. How do we structure the data? In particular, how do we discover the true causal relationships?
6. What is the best mathematical representation of these relationships? How do we estimate them?
7. How big is the causal effect of each variable? Which are the most effective variables to manipulate?
8. How do we select the right business interventions? What returns can we expect from investing in these interventions?

be poorly defined and of dubious causal value in driving business performance [Gupta et al. 2005].

A deep dive into the psychology literature that provides a foundation for these constructs reveal that they are usually produced through factor analysis, which suffers from the garbage-in garbage-out problem, i.e., the factors owe their existence to a set of measurement items constructed by a researcher with a particular personal interest in those items, and there is no solid empirical grounding that supports the selection of those items or otherwise guides the formulation of those factors (Section 3.1.3.2.1). Invariably, the factors found correspond to theoretical preconceptions by the researchers; different preconceptions have led to different and confusing formulations. If we fail to correctly conceptualize the psychological constructs, we risk failing to acquire data on critical explanatory factors, or we risk studying variables that are only somewhat relevant but also contain a lot of noise that may weaken the explanatory power of the model (e.g., “values” instead of “emotions”).

Some solutions to these methodological challenges have begun to appear in the strongly empirical studies known as "second-generation cognitive science," specifically the work of Lakoff [1999], and Damasio [1994], and others [Schall 2001; Bechara et al. 1997, 2000] who rely on empirical research in neuroscience, including MRI studies, to identify the constructs actually represented and processed in the brain. Our study is among the first to attempt the application of these laboratory findings to the design of observational field studies. In particular, this assisted the creation and use of a `psychological metamodel' to guide our variable selection, data collection and modeling activities, and further influenced our choice of methods for statistical analysis, e.g., to rely more on tetrad-based causal discovery algorithms rather than factor analyses (Section 4.1).

Another methodological challenge in studying the customer arises from the temporally protracted and spatially distributed nature of the problem. The process of purchasing a TV spans many months; at the beginning of the study, the best information we could obtain from our client suggested that the decision process lasted about two to three months, with a certain proportion of "impulse buyers" completing their purchase in much less time. However, one of the subsequent findings from the study turned out to be that the process, from earliest awareness of need to placing the order, actually spans over two years! This makes it difficult to capture information from all parts of the process that could eventually affect the final purchase decision. Even when the time span is on the order of weeks, the methodological challenge is to capture the events, states and activities in customers' lives that are *relevant* to the ultimate purchasing model. The difficulty lies both in knowing what might be relevant---ahead of time, before we have built the model---and of capturing data, often repeatedly (e.g., from multiple store visits) over a protracted time span. Although it is possible to retrospectively interview shoppers about their past experiences, that obviously produces errors from memory issues, rationalization of behavior, and other kinds of biases. Further, temporal information is often crucial to resolving causality; for example, if a flyer was seen by the shopper *after* visiting a store, we can discount the flyer's effect in inducing the shopper to visit the store. However, disentangling temporal sequence is much harder in retrospective interviews.[2] Clearly, real-time data is strongly to be preferred, if it is at all possible to obtain it. Additional difficulties with interviews arise when the interviewee describes other actors, e.g., salespeople encountered during a shopping trip, since the shopper's attention was only partially on the salespeople (being more preoccupied with their own shopping) and therefore cannot provide a detailed description of what the salesperson did. There is a need to observe other environmental data that the shopper may never notice, e.g., that there were cheaper TVs in the next aisle, but because of their particular placement, the TVs were not noticed and thus the shopper got a misleading impression about the retailer's prices. There is a need for external observation, e.g., by means of video surveillance or by an ethnographer. All these challenges led to another significant characteristic of the methodology that we created for the study, namely the use and adaptation of structured observation methods from ethnography (Section 3.1).

[2] Peeking ahead at one of the significant findings from the study: it is nearly impossible to disentangle the sequence of online activities via interviews. Many people cannot even distinguish multiple `sessions' (multiple times they go online to a Web site); they remember multiple sessions as a single visit.

Assuming that we have succeeded in obtaining such a rich collection of qualitative data, via ethnographic, psychological, or other methods, the next challenge is how to analyze it in a rigorously scientific manner that ultimately enables us to explain the mechanisms that drive the outcome, and allows us to systematically derive business strategies. Unfortunately this has traditionally relied on the researcher's skill in interpreting the data to `diagnose' the issues and directly come up with business recommendations (e.g., see [Arnould and Wallendorf 1994]). While there is no doubt that `smart' researchers do come up with useful suggestions, it is far more common for the researcher to regurgitate conventional wisdom from whatever discipline or work experience they have previously embedded themselves. So the diagnosis usually turns out to be a "communication issue" if the researcher has a degree in communications, a "relationship problem" if the researcher's focus has been social relationships, a problem of "designing the right tools" if the area of interest is computer technology, and so on. Particularly good consultants diagnose the problem as whatever corresponds to their business sponsor's theory about the critical factors. Even when the recommendations are genuinely creative, there is suspicion, usually justified, about the ethnographer's or psychologist's qualifications to directly come up with business recommendations. Clearly there is a need for more science in the analysis.

While it is rare in business environments to do rigorous qualitative analysis, it more common for academic analyses to use more robust procedures such as having multiple researchers read or listen through the transcripts of the qualitative data, and interpret and code the data, and then to do comparative or inter-rater analyses to obtain a higher degree of reliability and validity. Unfortunately almost all qualitative analyses rely on the identification of "themes" as the intermediate step; such thematic analyses provide no systematic method for reasoning about how to affect the outcome of interest and create business recommendations. Once again it is back to skills of the researcher. Plausible qualitative theories may have been created, but there is no way test the theories in a robust manner before committing to a business decision. See Section 3.1.6 for an alternative approach to qualitative analysis.

In principle, any qualitative theory can be turned into a quantitative theory by obtaining large amounts of data on the qualitative variables and doing statistical analyses on these variables. In practice, this transition from qualitative to quantitative methods has traditionally been one of the biggest failure points of market research. The main reason is that qualitative researchers such as ethnographers do not have the conceptual vocabulary and training to produce the kind of information that a statistician needs downstream, let alone a business decision maker even further down.[3] "Themes" are not a good input to a statistical model. Nor are diagnoses of "relationship and communications issues". How does one turn material from interviews or a focus group into grist for a statistical model? While a qualitative researcher can invent specific questions for a survey based on their own interests and predilections, they have no means of figuring out what the design of a

[3] For two notable exceptions to this remark see the insightful papers by Katz [2001, 2002], and Chapter 6 in [Miles and Huberman 1994]. Unfortunately these writings have been largely ignored by the mainstream of ethnographic practice and are rarely utilized in ethnographic analyses. And even these writings, great as they are, have a long way to go towards supporting scientifically rigorous inferences.

specific question, or the omission of a question, would do to the statistical model. In particular, what are the relationships of the questions to each other and to the primary outcome? This lack of knowledge often drastically limits the scope and validity of the conclusions drawn from the survey, because of a number of issues such as inappropriate granularity, and statistical confounding. Conversely, it is common practice today for a psychometrician or statistician designing a survey to refuse to take responsibility for the substantive content of the study; after all, it is the qualitative researcher who spent most of the effort embedding themselves in the scene to gain an intuitive understanding of the problem, and is thus the "substantive expert". The result is that statisticians makes numerous assumptions about the substantive content of the model; assumptions that are not documented anywhere, not even qualitatively examined, and eventually swept under the rug. The price of course is that the familiar `garbage-in garbage-out' syndrome afflicts the quantitative model. Since the only mode of information transfer between the qualitative and quantitative researchers is conversation, it not a surprise that much of the scientific rigor that may have been originally present in each of the two components is completely lost in the transition from the former to the latter.

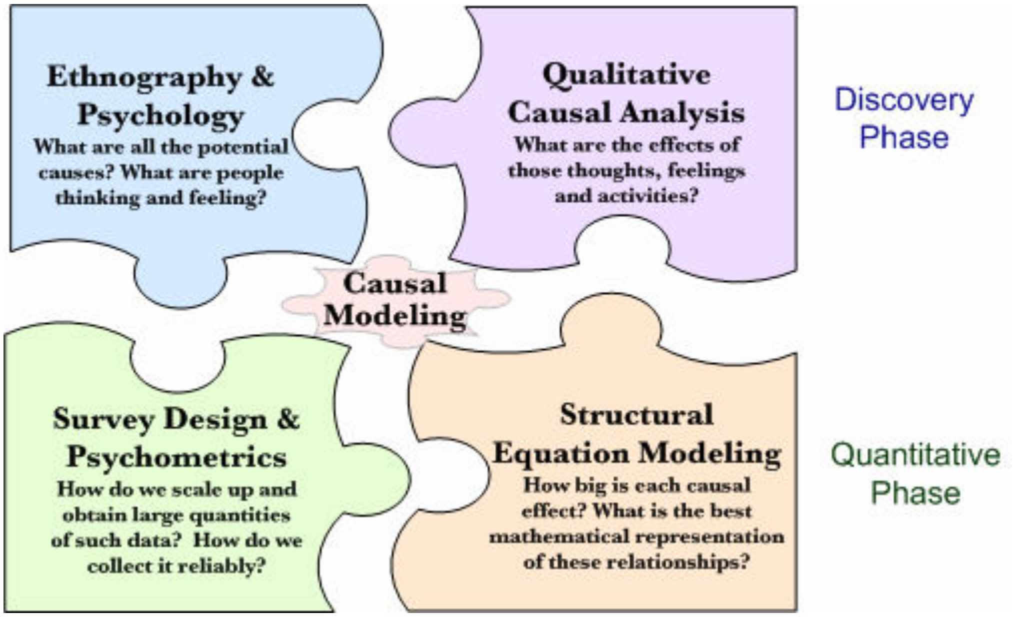

Ethnography & Psychology
What are all the potential causes? What are people thinking and feeling?
Qualitative Causal Analysis
What are the effects of those thoughts, feelings and activities?
Discovery Phase
Causal Modeling
Survey Design & Psychometrics
How do we scale up and obtain large quantities of such data? How do we collect it reliably?
Structural Equation Modeling
How big is each causal effect? What is the best mathematical representation of these relationships?
Quantitative Phase

Figure 4. A methodology for building causal models: The pieces of the puzzle.

One of the biggest insights that we realized during our research is that causal models are the solution to this disciplinary gap, because they have both rigorous qualitative and quantitative semantics, unlike any other analytic structure that we know. Ethnographers do not have to learn statistics to build qualitative causal models, although they do have to learn a new set of qualitative concepts.[4] Conversely, statisticians don't have to embed

[4] While qualitative research textbooks and other sources have for a long time described models of `cause and effect' [e.g., see Miles and Huberman 1994], those methods merely prescribe diagramming our

themselves in the field to obtain the substantive knowledge they need in order to resolve model structure; a rigorous qualitative causal model can be robustly transformed into a statistical model, without losing critical information along the way.[5] In other words, the formal semantics of causal models provide a lingua franca that enables the design of a methodology that restores scientific rigor to the transition between qualitative and quantitative research. Widespread adoption of formal causal modeling techniques offers the hope of a substantial jump in the quality of market research in general, and greater reliability in business decision making.

Having made the transition from qualitative to large scale quantitative data, numerous statistical methods present themselves as potentially providing insight into the data. Discrete choice models, hierarchical Bayesian models, and a number of regression techniques including structural equation modeling (SEM) stand out, as well as a slew of data mining techniques for creating exploratory models. While some of these techniques are complementary to or at least not inconsistent with our approach, there were several guiding principles that informed our choice of methodology. First of all, since our goal was to *affect* the outcome, preserving causal information was a priority. Therefore methods that played around with correlations or associations and required unjustified leaps of causal inference in order to derive business conclusions were deemed undesirable, which meant that many traditional data mining techniques were unusable (see Section 4.1.1). On the other hand, SEM as reconceptualized by Pearl [2000, Chapter 5] dropping its traditionally obscure terminology (e.g., with respect to exogeneity, identifiability, and computation of effect) in favor of more intuitive causal concepts, was a perfect tool. A second consideration was our preference for structure discovered from the data over structure imposed by an `expert'. Although we had the advantage of robust transfer of qualitative knowledge via a qualitative causal model, we preferred to use that knowledge only in places where automated modeling was inadequate, and substantive assumptions were unavoidable. Moreover, we needed automated help with structure discovery because the enormous size of the model precluded manual specification of every link. This necessitated the design of a quantitative analysis methodology that placed a heavy reliance on automated analysis tools, while providing a user-interaction facility that supported disciplined and systematic introduction of assumptions into the modeling process (Section 4.1). To support this methodology we developed our own quantitative analysis tools, the Causal Modeling Workbench, via a combination of off-the-shelf statistical software (SPSS, Mplus) and custom code built upon the R platform.

The wide-ranging nature of the challenges and solutions described above clearly calls for a new unified methodology that develops systematic interfaces between several disciplines, preserving scientific rigor when crossing disciplinary boundaries (Figure 4). While causal modeling concepts constitute an important part of the glue that ties these

everyday understanding of such relationships, e.g., via influence diagrams or fishbone diagrams. Such models lack important characteristics of causal models as described by Pearl [2000], e.g., qualitative concepts such as confounding, blocking, intervention, and interpretation of omitted links as independence, which carry information that is critical for subsequent quantitative analyses.

[5] Put another way, this is a major improvement with respect to the garbage-in garbage-out problem of statistical modeling.

pieces together, there is also a need for innovation within each of the disciplines, e.g., in migrating ethnographic observation methods from the study of culture to the study of causal mechanisms in the field, or in developing statistical modeling techniques that explicitly support insertion of qualitative substantive knowledge into automated discovery algorithms. We will further describe this methodology in Sections 3.1 and 4.1).

Based on the above methodological foundation, we summarize our framework for business strategy in Figure 5. A critical first step is the identification of the key business outcomes of interest. While this seems obvious, in practice this step is commonly prone to errors in the form of invalid hidden assumptions and improper operationalization. For example, “increased customer satisfaction” is a popular metric of success in many businesses. If we investigate why business managers use this metric, we usually uncover tacit or explicit beliefs that increased customer satisfaction raises revenues or creates loyalty, but further investigation yields no empirical evidence for these beliefs. If the underlying goal is to increase sales through customer satisfaction, it is better to specify sales as the primary outcome, and set aside customer satisfaction as an intermediate variable that may or may not enter the causal model, depending on what the discovery-phase methodology uncovers. If the modeling reveals that satisfaction is merely correlated with sales, it is of much greater business value to find the variables that causally affect sales; the fact that some of these variables also affect satisfaction would be an academic curiosity with no business actions implied. Thus, in general end-goals should be preferred over proxy goals, in order to avoid hidden assumptions. Further, the measurement of such goals must be unambiguous. Are we measuring satisfaction with a product, with the service, or with the shopping experience? The literature shows these are quite distinct, and the business interventions required will differ for each. Similarly, if “improved customer experience” is an end-goal, how do we operationalize the concept? Since product purchases are often family decisions, is it sufficient to measure a single person’s shopping experience? The solution to these difficulties often involves recognizing that many common goals are really intermediate outcomes assuming tacit theories should really be explicitly studied via the modeling process; these variables are not really ends in themselves. The preferred approach is to fall back on more tangible variables such as sales, profits, and so on, as primary outcomes. A key insight that we learnt from our study is “customer experience” and “satisfaction”, which are two of the most central concepts in interaction design and marketing, are broken to the point of being unusable for any practical modeling purpose.[6]

[6] “Customer experience” turned out to be broken concept because purchasing televisions turned out to involve individual and joint activities by a shopper, the shopper’s spouse, and their kids, and all of these had causal influence on the purchase decision. Because of the multiplicity of actors, and the causally interrelated nature of their shopping activities and events, there isn’t a single “experience” whose quality can be assessed. The nearest equivalent turned out to be the notion of a shopping “trip” to a store. This too was not usable as a proxy outcome, since the purchase outcome depended on multiple trips, and there does not appear to be a causally meaningful way of combining “trips” into an “experience” with a retailer whose goodness can be measured in a way that explains the purchase decision.

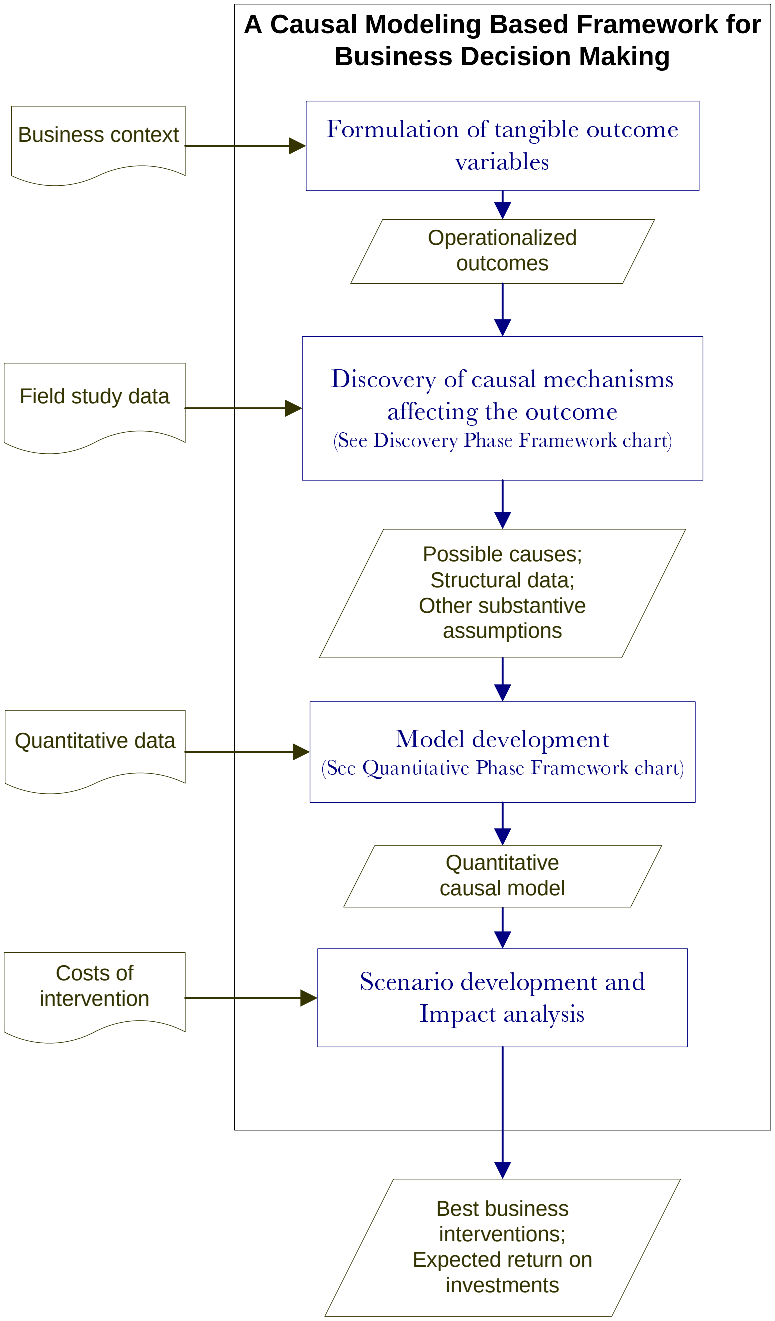

A Causal Modeling Based Framework for Business Decision Making
Business context
Formulation of tangible outcome variables
Operationalized outcomes
Field study data
Discovery of causal mechanisms affecting the outcome
(See Discovery Phase Framework chart)
Possible causes; Structural data; Other substantive assumptions
Quantitative data
Model development
(See Quantitative Phase Framework chart)
Quantitative causal model
Costs of intervention
Scenario development and Impact analysis
Best business interventions; Expected return on investments

Figure 5 Outline of our overall framework

Steps 2 and 3, namely the discovery and modeling of causal relationships are described at length in Sections 3 and 4. The final step, using a completed quantitative causal model to derive business strategies bears explanation here.

A causal model in essence provides "a look under the hood" of the business. It provides a picture of how the business is functioning, and what can be achieved by controlling each of the thousands of variables in the model. For each intervention made by the decision maker, changing any given variable (or combination of variables) by "1 unit change”, the causal model computes the resulting number of units gained on the primary outcome (e.g., sales). Thus, if the decision maker knows how much it will cost them to obtain a 1-unit change in a given variable, he or she can compare that against the predicted gain in sales, and infer the cost-worthiness of making that intervention. For example, if a $50 rebate on every TV in the store yields a 3% increase in TV sales, and the average price of a TV is $500, the rebate is not worthwhile. On the other hand, if a $1 million investment in retraining the salespeople can produce a ½-unit improvement (on the Likert scale) in the perceived helpfulness of these salespeople, and if that improvement produces a 10% increase in sales, the investment would be recovered over a full year’s sales of TVs assuming enough units are sold. In other words, a business strategist can play “what-if” scenarios, querying the model about the results of a hypothetical strategy or intervention.

When the intervention is specified in the form of a change imposed on variables that are explicitly present in the causal model, the model directly computes the resulting change in outcome, using the standard mathematics of effect computation in causal graphs. On the other hand, if a creative decision maker constructs a hypothetical intervention that involves variables that are not present in the causal model, two approaches can be used. The simpler approach is to speculate about the effect of such an intervention on variables that are already present in the model; this is usually quite easy, because the process of coming up with the creative intervention involves hypotheses about why the intervention would work. Following the intuition that suggested the intervention quickly results in pointing to a few variables in the causal model that would be affected the intervention, and from that point onwards, the causal model predicts the results. So if the decision maker hazards a guess, typically a range of values, about the effect that their creative strategy would have on the variables in the causal model, a resulting range of results can be computed. As in most what-if analyses, this quickly produces a rough feel for the effectiveness or ineffectiveness of the strategy, which is the primary value to the decision maker.

The second, somewhat more complex but more useful, approach is to work backwards from the outcome towards the variables which are being considered as the subject of a business intervention. Reversing the mathematics of the causal model, it is possible to compute the size of intervention required to obtain a given improvement in the outcome. For example, it may turn out that a 0.12-unit change in “perceived helpfulness of salespeople” may be required to cause a $1 million increase in revenue. This automatically defines a break-even point which implies that any intervention such as training salespeople would not be worthwhile if it costs more than $1 million to achieve an observable gain of 0.12 units on a helpfulness survey. Tests can then be devised to reduce the decision-maker’s uncertainty about what results a particular training intervention would produce. Thus the causal model provides quantitative bounds that help assess the value of a strategic intervention, even when precise effects are unknown.

In our experience, the application of causal modeling that is most popular with business decision makers is the prioritization of strategic initiatives. While a full-blown causal model reflects the complexity of the business through thousands of variables, many business decision makers often want this to be reduced to a simple picture: “What are the Top Ten critical success factors that I should focus on to turn my company around?” In the presence of cost information, this is easy to answer. The advantage of the causal model is that it computes the *total* effect of any intervention: it propagates the change over all possible paths (causal mechanisms) from the intervention to the outcome. Therefore superimposing the cost of intervening on each variable in the causal model allows us to formulate a straightforward optimization problem: identify the intervention that maximizes return on investment, i.e., gain in the outcome variable divided by the cost of the intervention variable. In truth, this can get complicated since a change in one variable may produce a change in several others en route to the final outcome, and therefore a simultaneous intervention on a group of variables cannot be simply computed by adding up the gains from the individual interventions; a more complicated causal calculus must be used. However, for practical purposes it often suffices to consider intervening on one variable at a time. Then the formulated optimization easily produces a sequence of the top ten variables that are the best business investments. The resulting simplicity appears to be psychologically important to business managers, who need to explain and communicate their priorities and investment decisions to their organizations and shareholders. When supplemented with a simplified causal diagram comprising just these critical success factors, it proves a lucid basis for rationally explaining their decisions, as well as a basis for identifying central themes around which top managers can construct their “vision”, “mission statement”, and so on. Thus causal models provide a solid, defensible, scientific foundation for constructing good business strategies.

3 The Discovery Phase

3.1 Study design and methodology for the Discovery Phase

The Discovery Phase of our modeling framework is illustrated in Figure 6. In a nutshell, the primary function of this phase is to resolve the garbage-in garbage-out problem (described in Section 2) that would otherwise afflict the Quantitative Modeling phase. In other words, the Discovery phase ensures that the right outcomes are being studied, that an adequate set of possible explanations have been brought into consideration, and that the numerous hidden assumptions that lie underneath quantitative models have been made explicit and either discarded or grounded in solid empirical data. Therefore this part of our framework is designed to produce good answers to the first three questions in Figure 3.

For example, the issue of model completeness is addressed here, i.e., how do we ensure that we have as complete a list as possible, of the variables that might turn out to be important influencers of the outcome under study? It is often easy to guess some of the causal variables, but how do we gain some assurance that critical variables have not been omitted? “I know advertising drives some customers to my store; but what are all the other reasons people are visiting us?” “My store carries every type of TV; but how come customers say they are still unable to find the right TV?” A missed “confounding” factor disrupts our ability to pin down causality, as we explain in Section 3.1.2. Finding the right variables becomes harder when customers’ inner thoughts and feelings are involved. “What are shoppers thinking when they enter my store?” “Why do they perceive me as expensive when my products are priced similarly to my competitors'?" The Discovery Phase answers such questions.

Note that the Discovery Phase may or may not be an end in itself. That is, we could stop at the end of this phase and proceed to design business interventions based on the qualitative insights gained. However there is often a lot to be gained by going on to build a quantitative model, such as understanding scale and priorities. When quantitative models are needed, it becomes important for the Discovery Phase to produce output of a kind suitable for rigorous input into a quantitative methodology. This affects how we design the both the data collection process and the analysis methods used in the Discovery Phase. There are many qualitative techniques that can provide useful insight of one type or another, but few qualitative techniques are designed to reliably transfer these insights into subsequent quantitative modeling efforts. In our framework, the most important output of the Discovery Phase comprises qualitative causal graphs, which follow the semantics used by Pearl [2000].

We do not actually need to produce a completely specified causal graph in which the presence or absence of causal relationships between every pair of variables has been determined. Indeed when hundreds of variables are present, there are thousands of links to be specified, which makes it difficult to manually construct these graphs through qualitative analysis. The difficulty arises in part because in a properly specified causal graph the *absence* of a link contains extremely important information, namely independence. Therefore we would have to examine every pair of variables in the graph to ensure that we haven't unintentionally asserted independence by forgetting to add a link. This kind of pairwise comparison is tedious and unnecessary for the needs of the Quantitative Phase. The minimum that we do need is (i) a specification of the variables that could directly or indirectly act as possible causes of the given outcomes, and (ii) "first-class causal mechanisms", i.e., a set of links and qualitative functional descriptions that specify which of the variables have been *observed* during the discovery phase to causally affect other variables. These mechanisms have been termed "first-class" to evoke their grounding in empirical data and our resulting confidence that such mechanisms exist in nature. In general it is theoretically possible that any variable could influence any other variable, but if we have not observed empirical evidence for it, and merely hypothesize the existence of such mechanisms, such relationships would not be called `first-class', and we would associate lower levels of belief in these assumptions during the quantitative phase. Identifying first-class causal mechanisms provides us with a useful shorthand notation that tells us which relationships must indeed be explicitly accounted for during quantitative modeling, which relationships can be tacitly ignored unless flagged by the quantitative algorithms, and how to prioritize the trustworthiness of any substantive assertions that we inject into the model (Section 4.1.2.3). In small causal models (say a few dozen variables) it is feasible to go all the way and completely specify a causal graph through manual qualitative analysis. Such completely specified graphs have extra value in the sense that they can be used as end-product in itself at the end of the Discovery Phase: they can be used qualitative reasoning and scenario planning, instead of for quantitative modeling. This has important applications in forecasting and long-range decision-making.

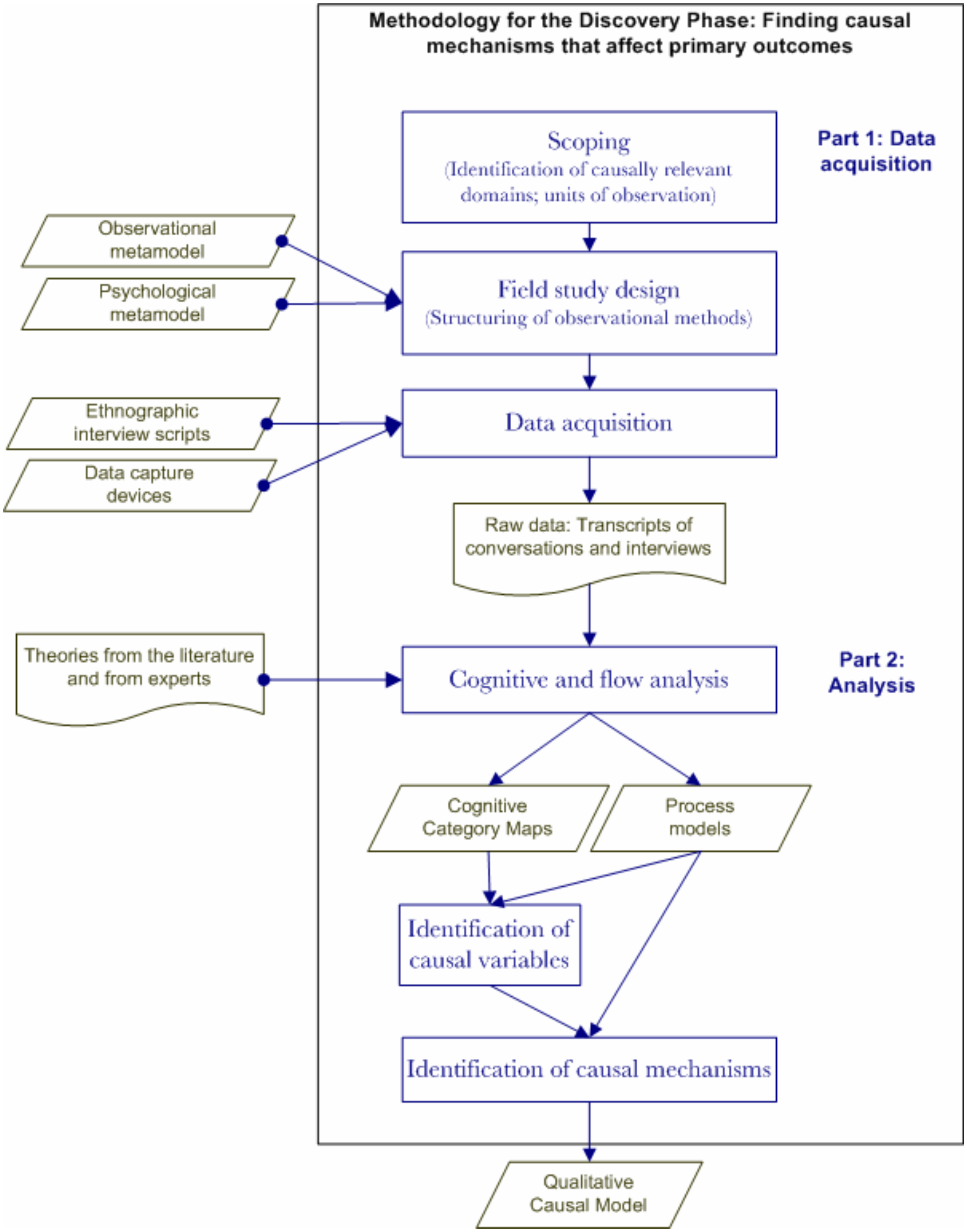

Methodology for the Discovery Phase: Finding causal mechanisms that affect primary outcomes
Scoping
(Identification of causally relevant domains; units of observation)
Part 1: Data acquisition
Observational metamodel
Psychological metamodel
Field study design
(Structuring of observational methods)
Ethnographic interview scripts
Data capture devices
Data acquisition
Raw data: Transcripts of conversations and interviews
Theories from the literature and from experts
Cognitive and flow analysis
Part 2: Analysis
Cognitive Category Maps
Process models
Identification of causal variables
Identification of causal mechanisms
Qualitative Causal Model

Figure 6. The Discovery Phase of our overall framework
(corresponding to Step 2 in Figure 5)

Figure 6 outlines some of the central activities, tools, and work products of the Discovery Phase. While some of these are new, others are derived from the usual best practices in qualitative research, and design of this phase reflects our emphasis on obtaining high-quality data of a kind that helps tease out the causal mechanisms at play. Many of the design constraints that resulted in this choice of methodology are brought out in the following section.

3.1.1 A theory-less discovery-oriented approach

Our earlier discussion of `first-class' causal mechanisms exemplified a more general principle underlying the design of our framework: empirical data should play a much greater role than preexisting theories and hypotheses in determining the structure of our model. Our ideal scenario is to work from a clean slate: the Discovery Phase starts with nothing more than a specification of the outcome variables that are the focus of study (e.g., sales), and proceeds to acquire the right kind of empirical data, and do the right kinds of analysis to construct a model that explains the workings of the given domain. In particular it does not rely on folk theories of decision making, e.g., "people become aware of a need, search for information, evaluate alternatives, compare alternatives, and select the best one" [Lilien et al. 1992, Kotler 2000, Butler and Peppard 1998]. Nor does it use popular hypotheses (e.g., "prior purchase causes satisfaction, which causes loyalty, which causes repeat visit and subsequent purchase") to set up models that must subsequently proven to be true or false (see Section 4.1.1 for a discussion of the confirmatory structural equation modeling approach). In other words, our framework replaces the usual 'hypothesis-driven approach' with a more `discovery-oriented approach'.

The benefits of this approach are reinforced when we empirically observe behaviors that differ from what was anticipated by marketers and other domain experts. For example, it turns out that decision making is not a process of comparison and tradeoffs between the available choices; shoppers do not compare all the TVs in a store, e.g., because they simply don't notice that there is a whole row of TVs on the other side of the aisle. Shoppers violate common axioms of rationality such as the independence of irrelevant alternatives. They like a particular TV because of the nice stand that it is sitting on, but do not trade off the stand against other TV attributes, contradicting common assumptions about consumer choice behavior. To increase the likelihood of discovering true patterns of behavior, it is therefore important to adopt a research philosophy of simply describing things as they are, minimizing the use of *a priori* theories. Later analysis of this data will uncover the structure needed for theory formulation.

Our approach is philosophically similar to that of grounded theory which is also an emergent-research method. However the analysis process is quite different, since our primary objective is to uncover causal mechanisms that are likely to hold in larger populations than the cases studied; our goal is not to rationalize the particular cases observed. Thus accurate description is important to our approach and is not an incidental means for generating concepts as in Glaser's approach; we do aim for the "truth", not just a conceptualization [Wikipedia: Grounded theory (Glaser)]. We then construct process descriptions as an intermediate analysis step before the main step of identifying causal mechanisms. Mechanisms are identified by starting from the primary outcome of interest and chaining backwards. This backward search for explanations guides the selection of variables or `categories'; categorization is not a goal in itself (other than what we need for cognitive category analysis). Our methodology does not look for "themes" and therefore activities such as memoing are not central. Glaser & Strauss' "theoretical sampling" is acceptable only when a quantitative model will be built; to be avoided in a qualitative-only approach because of bias. Comparisons are useful assessing the presence of causal links and when making generalizations, but otherwise not central to our methodology. There are many versions of grounded theory in practice, and our methodology is completely unrelated to variants such as ethnomethodology [Crabtree et al. 2000]. The latter also rejects a theory-based approach, but provides no guidance on what to do instead; the researcher is left helpless. We rely on structured methods such as the use of metamodels and causal analysis; indeed we believe that one of the measures of the effectiveness of our methodology should be that we have provided enough guidance on technique to ensure that multiple researchers can independently draw the same inferences from a given dataset.

While the value of a theory-less approach is obvious in terms of providing true insight and avoiding regurgitation of analysts' favorite theories, we are not arguing that input from subject matter experts should be totally ignored. We merely argue that it a much lower value ("belief level") should be placed on such input, relative to direct empirical observation, when injected into the model, and extra care is required during qualitative analysis to avoid being subtly biased by these theories. (See, for example, the principles articulated in Section 4.1.2.3.) Despite this, there are at least four benefits from including expert input and theories from the literature into Discovery research. An obvious benefit is that adds a greater level of confidence in the completeness of our model, i.e., that we have ensured the inclusion of all variables that might potentially have a significant causal effect on the outcome. Secondly, disconfirmation of preexisting theories can be a great source of insight. Thirdly, in business consulting settings there are solid political reasons for inclusion of such input: the business decision maker may dismiss the results of a study if their pet theory was not confirmed or disconfirmed by the study, since their alternative theory still stands. And lastly, it is a lot cheaper to ask somebody than to conduct a field study; the money and time available for conducting discovery phase research induces a tradeoff against model validity. In short, our framework in Figure 6 is not meant to imply a single flow of activities; rather, it permits pragmatic compromises between grounding in high-quality empirical data and shortcuts using experts' theories, giving preference to the former where possible.

It is important to note that our Quantitative Phase methodology substantially mitigates the biasing effects of these prior theories, for two major reasons. First, the Discovery Phase only needs to produce the *possible* causes and causal mechanisms, not the definitive causes and mechanisms. The quantitative phase will weed out mechanisms that occur infrequently or never, and therefore including variables that are merely hypothesized to be causes does not hurt much.[7] The second form of protection arises because our quantitative phase utilizes a unique causal discovery methodology that initially ignores all the information produced during the discovery phase (except for the choice of variables themselves), and uses special algorithms to uncover the relationships between variables (see Phase 1 in Figure 19). Substantive theory is subsequently injected

[7] The main disadvantage is that it increases the size of the model and adds work (since modeling effort goes up in proportion to n^2 where n is the number of variables). However the extra work is often worthwhile because the extra variables provide a safeguard against confounding and identifiability issues.

iteratively; in other words, theoretical assumptions are applied only to links where causal directionality could not be automatically inferred. Furthermore, the algorithms are designed to minimize the number of theoretical assertions used, and to make these assumptions in decreasing order of trustworthiness; see the design principles in Section 4.1.2.3. Therefore much of the risk from using these theories is reduced. The main negative effects that remain from using these prior theories arises from (a) the choice of variables themselves: if reliance on poor theories makes us omit key variables, we will find out later that the variance of the outcome variable is insufficiently explained, but we will not be able to fix the omissions; indeed we increase the risk of model error (b) subtle reliance on these theories during Discovery Phase analysis when we identify causal variables and mechanisms (the lower half of Figure 6); and (c) failure to discover new functional forms (which explain how the value of one variable is determined by the values of its parents), e.g., see our discussion of quasi-deterministic models in Section 3.2.3 and nonlinearities in Section 4.3.6.

3.1.2 Exploring the domain: scoping

A theory-less approach poses significant difficulties in data collection and analysis: How do we uncover the variables that could possibly influence the primary outcomes? Where do we begin our search, and how do we know we have not missed important variables?

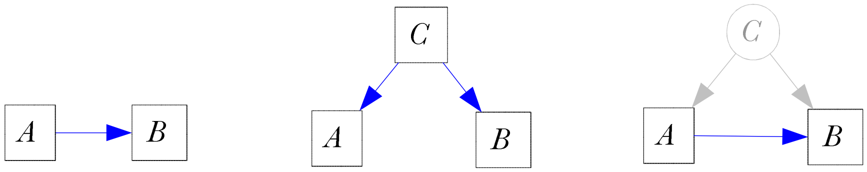
It is important to note that at this stage of the game, being able to actually acquire the data is *not* a constraint. It is sufficient that on paper we identify all possible factors that *could* cause changes in the outcomes. Variables that later turn out to be difficult or impossible to observe may be treated as latent variables [Bollen 2002].   The primary activity at this stage is ensuring that we find the variables that could be relevant. Of course it is in general impossible to prove that we haven’t missed a variable since there may be hundreds of factors that affect an outcome. The practical goal is then to perform due diligence in finding a set of variables that we believe explains “most” of the behavior of the outcome, leaving behind (a possibly large set of) variables whose effects would be considered a small and acceptable level of “error” in our explanations.

The principle of locality (“no action at a distance”) [Pearl 2000] provides the basis for our search.   Rather than try to list all possible direct and indirect causes of the outcome variable of which there are infinitely many, we search only for the “direct” causes, i.e., a small set of variables that are in some sense immediately adjacent and local to the outcome, usually physically adjacent.   So for example, for our primary outcome which is the decision to purchase a TV from Retailer A, the immediately adjacent variables belong to the final shopping trip made by the customer, e.g., the last few TVs considered, the salespeople’s behavior on this particular trip, and so on. Earlier events, such as seeing an advertisement that induced the shopper to visit the particular store, are ignored; these are not direct causes of *Purchased*; rather they are direct causes of the intermediate variable *Visited the store*, and thus indirect causes of *Purchased*.   A series of backward chaining observations (What happened in the final minutes before purchase? What happened in the few minutes before that… what happened before visiting the store… etc.) leads to the scenes where indirect causes were at play.   Systematic tracing of this kind, focusing on local events in each step, leads to the systematic uncovering of all the observation scenes in which we are likely to find variables that could causally affect the outcome.

It is trickier to apply the principle of locality to the human mind: in the final moments before purchase, the shopper may have remembered events from long ago that significantly influenced their decision to buy. While the principle still applies---only the memories that are recalled would relevant as direct causes---it is much harder to observe the "scene" in the shopper's mind, which is necessary for tracing back to the prior relevant events. With current `mind-reading' technology such as MRIs and galvanic skin sensors being ineffective outside the laboratory, we had to rely on verbal methods such as the ThinkAloud protocol for eliciting this data. Put another way, if the causal variables involve humans, our traceback procedure requires identification of all causally relevant `actors' and (verbal) elicitation of their mental processes in addition to behavior variables acquired by direct observation of the physical scene.

While backward tracing of observation scenes is simple in concept, it can be tricky in practice, and getting it right is much more important than it may initially seem. Its importance arises when we deal with the issue of confounding. To assert that *A* causes *B*, we have to be reasonably sure that we have not omitted a common-cause *C* that simultaneously influences both *A* and *B*, thereby causing the latter two to falsely appear correlated (Figure 7). A failure to include *C* in the model prevents us from correctly estimating the causal effect of *A* on *B*.[8] In observational studies, since we don't have randomized experimental controls, we cannot afford to miss variables that are indirect common causes. For example, if we study the effect of *Advertising flyers* on *Purchase*, we cannot miss the fact that the advertised products may also be placed *On sale within the store*; customers who have never seen a flyer may buy the advertised products because they walk into a store and find the products on sale. In other words, we need to observe both the in-store scene as well as the advertising scene, and if we omit the former, we would wrongly infer that seeing a flyer is responsible for the purchases. The more complex the domain, the trickier this gets; for example, when we study what makes a patient use *Prescription versus generic* drugs, we may be quick to observe the patient's world, and the doctor's world, but we may forget that the patient's medical insurance benefits were determined by the patient's employer, and the choice of drug was therefore heavily affected by the scene in which the employer's pharmacy benefit manager negotiated the insurance plan. One of the challenges with causal modeling in uncontrolled observational environments is that there is no way to *a priori* limit the set of observation scenes; the researcher has to go wherever a careful application of the traceback procedure takes him or her.

[8] This is an oversimplification; it is possible to correct for the omission of *C* if other "blocking" variables or "instrumental" variables are present. However, the current point still holds---we need the presence of the right variables in the model to permit the computation of the effect of *A* on *B*.



(a) A truly causes B.

(b) Common-cause C simultaneously affects both A and B, making them correlated even though A does not cause B.

(c) Correlation due to an omitted common-cause C is mistaken for a true causal link between A and B. C is a "confounding variable".

Figure 7 The concept of confounding

After we obtain a rough understanding of the "observation scenes" or domains that are causally relevant to our primary outcome, we need to make sure that our "units of observation" in those domains are also at the correct granularity to obtain the right data. This is analogous to how we choose the "unit of analysis" and "unit of coding" in qualitative research [Boyatzis 1998]. In Section 2 we explained how we originally had defined our primary outcome as the customer's "satisfaction with the purchasing experience", then came to realize that both "satisfaction" and "experience" were broken constructs, and ultimately settled on "decision to purchase from Retailer A" as a much better defined outcome that avoided many hidden assumptions. We did continue to harbor certain assumptions that were the original motivators of the study, e.g.:

> *H1: The events that occur during a purchasing experience and the (possibly intangible) qualities of the purchasing experience (such as store atmosphere) affect the choice of retailer.*

It is good to explicitly acknowledge these hypotheses, not just because they are often the *raison d'etre* of the study, but because calling them out makes it easier to point out factors not stated in the hypothesis (e.g., the customer's personality) that may affect the outcome, and help to frame our discovery phase methodology as a non-experimental approach which seeks not to validate or disprove a hypothesis, but to find previously unknown and insightful factors that are important for controlling the outcome. These other factors in turn refine our choice of units of analysis and observation. For example, based on the above hypothesis, we initially defined our unit of analysis as the individual purchasing experience, starting from the moment when the shopper initially becomes aware of the need to buy a TV, to the point where they purchase a TV or decide not to buy a TV.[9] However it became clear that more than one shopper is likely to participate in "the experience" and there is no logical basis for taking e.g., the husband's view as being more "primary" than the wife's, for the purpose of causally explaining the

[9] Or fail to buy a TV within a given amount of time. This last situation is not uncommon as we discovered during the study, because TV purchasing is a protracted process sometimes lasting several years, and shoppers often temporarily put off the purchase because of other priorities, with no clear end in sight.

outcome. So the units of analysis and observation were refined to include the experiences of each of the "actors" that we discovered (see the discussion of the metamodel in Section 3.1.3). Likewise, there was no initial basis for determining what size of time slice we should observe in the scene, i.e., should an "episode" span minutes, weeks, or months? However early ethnographic interviews revealed the existence of the "shopping trip" as a salient concept, and therefore trips and visits to particular stores became units of analysis.

In general, taking a theory-less approach necessitates an iterative approach to choosing units of analysis and observation, where we analyze our initial data for causally relevant variables, and guided by that look for more detail that explains those variables in turn. Because hypotheses about causality can only be generated by working backwards from the outcome, it follows that there is no way to define *a priori* the units of observation and analysis for the entire study; that can only be figured out once we get some idea about the nature of the variables that might explain the outcome[10] . At the beginning, when we have no guidance at all, our metamodel is particularly useful to prime the pump. Once the specific actors and processes involved become clearer, we redefine the units of observation to zoom in for a closer look at the newly defined intermediate outcomes and their causes in turn.

3.1.3 Exploring the domain: using the metamodel

Even when we have some idea of the domains that may be causally relevant to the study, the problem of ensuring that our exploration of the domains is thorough enough to catch all the key variables requires systematic support. The problem surfaces acutely when we conduct ethnographic field trips, since there is a barrage of sensory data that must be captured and analyzed. Not surprisingly, field researchers need the help of some kind of checklist of what to observe and record. What makes this somewhat tricky is that a checklist that is too domain-specific can introduce subtle biases. For example a focus on conversational dynamics may generate variables describing the approachability of a salesperson as perceived by the shopper, but may omit the fact that the conversations themselves were significantly affected by the amount of traffic in the store and the number of salespeople available. Also, variables that are meaningful in one domain, such as within the store, may not exist in another domain such as the Internet; e.g., consider, looking up a product on a search engine. Generalizing a checklist to work across most domains has to be carefully done to avoid making the items too generic and then no longer to be useful in the field, e.g., "record all searches for information".

Many checklists have been developed in the ethnography literature, notably [Bell and Teague 2001, Schwartzman 1993, and Spradley 1980 p.82], and they have their respective merits and biases. Bell's checklist is the most detailed and useful for field use, although it omits some key areas e.g., from psychology. Spradley's conceptual organization is by far the best one, which is not surprising since his analytical methods (domain, taxonomic, and componential analyses) are also the most advanced. However Spradley's list is not sufficiently detailed for field use, and is also quite weak on the

[10] Of course it *is* possible to work out *a priori*¸ the units of observation and analysis for the outcome constructs, but that is only a tiny part of the problem.

psychological constructs. Looking in a completely different direction, we discovered that Spradley's conceptualization is in some ways remarkably similar to the `metamodel' underlying the Unified Modeling Language (UML) used by software engineers [Booch et al. 1999]. A "metamodel" is specification of the modeling constructs provided by a language. In hindsight, it is not surprising that a modern modeling language such as UML, which is designed to support description of business domains (although primarily as a means to identify software requirements), should contain some of the most important concepts studied by ethnographers.

Integrating all of the above sources, we constructed our own version of an ethnographer's aid or checklist. We call this an ethnographic metamodel, and use related terminology such as "metaclasses" (the categories listed) and "metarelationships" (general relationships between metaclasses). Although such terminology may be currently unfamiliar to qualitative researchers, it increases clarity in several ways. First it reinforces the point that unlike most `models' of a domain, our list is not domain-specific, and the specific instances of the metaclasses in our list will be identified by the researcher via field observation. This is extremely important in any emergent-research or discovery-oriented approach, which raises the question of whether *a priori* theories might creep in via our checklists and thus bias the study or turn it into a hypothesis-driven study. In response, note that theories generally specify models, not metamodels. Secondly we not only specify categories for observation, we also specify relationships among the metaclasses. For example, "emotions" are a type of (subclass of) "state". There is a lot of such domain-independent structure which is extremely useful in adding clarity to the modeling process, and simplifying analysis.

3.1.3.1 Overview of the metamodel

The metamodel specifies the kinds of things to be observed (metaclasses) and the relationships between those things (metarelationships). While the list presented here is the most complete general use list that we could construct for general use, each time the list is used in a new study, additional items specific to the study might need to be added to the list. We have grouped metaclasses into categories such as "Overall experience", "Actors", etc., for convenience since this a long list; there is no particular structure implied by these groupings. The categories themselves fall into two broad groups overall: *Basic categories* and *Special categories*. Basic categories are those that appear to be truly universal, and must invariably be studied for thoroughness. Examples are Actors, Objects, Acts, Spaces, Events, and Time. Special categories are described using two or more Basic categories in combination. Examples are Information, Traffic, Tools, Food, and Shopping. Special categories are not universal, but nonetheless have appeared often enough in the literature to merit being collected here. Drawing a lot of variables from any one of those categories, e.g., "information and communication" would give a particular slant to the study, which may or may not be considered a bias. We have also separated out the psychological variables because their origins lie in a completely different source: neuroscience rather than ethnography or IT modeling; however they are not fundamentally different from the other basic categories.

The items in each category are distinguished as either *participant* questions, *observer* questions, or both. Participant questions can be directly addressed to the actors in the scene; these questions can be the basis for interviews, for self-recording by the participants, and so on. Observer questions are more suited for unobtrusive observation, and often provide more objective but less informative data.

The sequencing of questions in each category is designed to elicit data in order of logical dependencies and decreasing order of reliability. It begins by identifying the actors and objects relevant to the category (i.e., “who/what”), and then asks “where” (location questions), “when” (time and event questions), “how” (method and process questions), and “why” (attitude questions). It is interesting to note that the first four questions appear to dominate the non-psychological categories, but the question “why” dominates the psychological categories. Each “why” question posed to a participant elicits beliefs, emotions, goals, and evaluations with respect to different things (cognitive categories), and therefore must be recorded along with the categories to which they refer.

It is important to note that the variables that will be ultimately extracted to build causal models are likely to belong to only to two basic categories: Acts/activities/behaviors, and Events/states. This is because of the nature of a “variable”: it must have the ability to take multiple values, such as doing/not-doing an activity, or being/not-being in a state. Actors, objects, places, etc., cannot be variables; only their attributes (which are typically states) can be variables. Nonetheless it is important to thoroughly elicit data on all the basic categories such as actors and objects, because they are the referents of the variables; missing a key actor will result in missing some key variables.

3.1.3.2 Basic categories

The overall experience

Both participant and observer:

- What is it like to be there?
- What is the first thing you notice? What is the last?
- What makes this place different from the office? Your home?
    - Analysis only: What makes this place different from other stores, from the Web?
- What are the rituals?
    - How do we detect this? Possible definition: Rituals are recurring activities done for reasons other than their functional purpose.

Actors / People

Participant:

- Who else is with you (relationship)?
- What is their age range?
- What are they doing?
- What objects are they using?
- What did they have with them?
- Where do actors place themselves? How do they use space? How are they moving around?
- Why are they with you?

Observer:

- Who else is in the space? Men? Women? Older? Younger? Ethnicity? Nationality? How are they dressed? How are they different from the staff?
- What kinds of groups are they in (families, tours, etc)?
- How long do they seem to stay?

Objects / Possessions

Participant:

- What are you carrying with you? Why?

Observer:

- Can you describe in detail all the objects?
- What are people carrying around with them (including purses, backpacks, mobile devices, etc)? How big are they?
- Where do people put them, keep them?
- How often do they use those things?
- What are people acquiring? What are they doing with it?
- Special categories of objects of interest to us?
    - E.g., products?

Acts / Activities / Behavior

Acts, Activities, Behavior, and Events are quite similar entities, and the following table is useful to distinguish their various connotations. Note that these differences do have modeling implications; e.g., Volition implies the presence of an Actor

	Volition (Intent/Control)	Atomicity (Short time span)
Act	✓	✓
Event	✗	✓
Activity	✓	✗
Behavior	✗✓	✗✓

Participant:

- What did you do?
- How did you do that?

Observer:

- What are people doing?
- What are other `actors' doing? (E.g., systems)
- How are they doing that?
- With what (objects) are they doing it?
- How do they vary over time? (E.g., what is different if you come back another time?)
- What are the ways acts/activities/events are related to intent/goals?

Events and States

Observer:

- What is happening to the actors?

Both:

- What happened? What was the situation before? What was the result?
- How do actors react?

Space / Place / The built environment

Participant:

- What does the space look like?
- How is it organized, partitioned, divided up?
- Is there a theme?

Observer:

- What does the space look like? How big is it?
- How is it organized, partitioned, divided up? How is it furnished?
- Where are objects located? How is space organized by objects?
- How is the flow of traffic organized? Structured?
- How is it signed?
- Is there a theme? (How is it supported?) Voice?

Time / Process

Participant:

- When did it occur?

Observer:

- What are the time/space cues?
- When does each `episode' (e.g., a session of use of the space) begin? End? How? Why?
- What marks the beginning and end? Is there a continuation? What marks it?
- What indicates or establishes the significance or meaning of an episode or event? (E.g., comments such as "This is going to be fun!")
- What are the relationships between shopping episodes and other life events? Do they occur in cycles or patterns?

3.1.3.2.1 Basic psychological categories

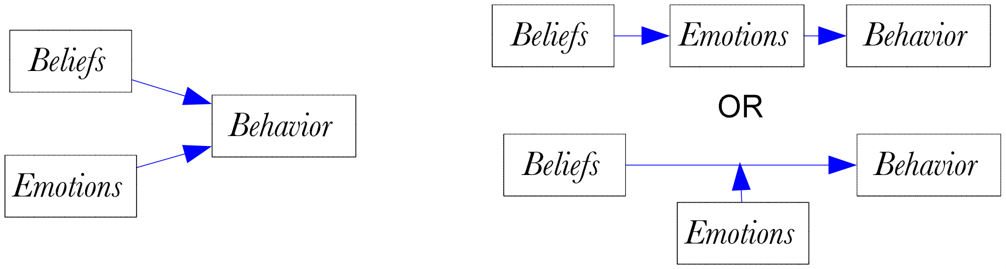
The traditional psychology literature and most consumer behavior studies grounded in that literature make heavy use of constructs such as "attitudes", "values", "affect" and so on as the hypothesized causal variables that influence observed behavior [Ajzen 1991]. For details on of some of major constructs traditionally used in marketing, see:

Attitude	Zanna and Rempel 1988, Cohen and Areni 1991, Worchel 2000, Petty et al., 1997, Ostrom 1989, Eagly & Chaiken 1993, Mowen & Minor 1998, Pratkanis, Breckler, et al 1989, Ajzen 2001, Limayem, Khalifa, et al. 2000, Jacoby et al. 1998
Affect	Erevelles 1998, Cohen and Areni 1991, Schwarz and Clore 1996, Scriven 2002, Swinyard 1993, Richins 1997, Batra & Ahtola 1990, Allen, Machleit, et al. 1992
Belief	Simpson, Weiner, et al. 1998, Tesser & Martin 1996, Ajzen 2001,

	Mowen & Minor 1998
Value	Scriven 2002, Tesser & Martin 1996, Kahneman 2000
Satisfaction	Giese & Cote 2000, Padilla 1996, Westbrook 1980, Mano and Oliver 1993, Shankar, Smith, et al. 2000, Gardial, Clemons, et al. 1994, LaLomia & Sidowski 1990, Ives, Olson, et al. 1983, Rushinek & Rushinek 1986, Oliver 1993, Wirtz and Lee 2003
Behavior	Donovan and Rossiter 1982
Flow	Novak, Hoffman, et al. 2000, Novak, Hoffman, et al. 2001

Unfortunately, close examination of the source of these constructs suggests that there was no real empirical basis for formulating human mental processing in those terms. For example, what is the empirical evidence that a shopper's "escapism" affects their "attitude" and causes a purchase behavior [Kim and Kim 2005], or that shoppers carry such attitudes in the first place at all? If we do accept the folk theory that "affect" plays a role in determining behavior, what precisely is that role, how does it play out in a given situation, and how does it relate to other psychological constructs such as "beliefs"? E.g., are beliefs and emotions additive, i.e., do they separately influence behavior and combine additively when both are present? A review of the literature reveals that these theories have been constructed by (a) some researcher inventing and operationalizing these constructs, often by systematizing folk theories, (b) conducting surveys to obtain data on the constructs, and (c) utilizing often suspect analytical procedures such as factor analysis (see our critique in Section 4.1.1.2) to `validate' the theory. Since a superficial fit with the data is usually found (as with most folk theories), that is taken as evidence that the human brain indeed processes "attitudes" and "values," even though the explanatory power of these variables in explaining human behavior is quite poor [Armitage and Connor 2001].

However a new line of research has emerged during the last decade, which utilizes MRI scans and other techniques to understand the workings of the brain, and in particular of the "mind" [Schall 2001, Damasio 1994, Lakoff 1999, Adolphs 2002, Bechara et al. 1997, 2000, Rees 2002, Rilling 2002] This work, called "second-generation cognitive science" has resulted in the identification of the mental constructs for which there is firm empirical evidence of their existence and properties. Notable among them are "emotions", "cognitive categories", and "images"; attitudes and values are notably missing and supplanted by the other categories. The interrelationships between all the constructs are not fully known, although it is clear that emotions mediate behavior, i.e., beliefs, memories, etc., all trigger emotions en route to triggering behavior. Knowing the existence of such relationships is critical to knowing how to generate good causal models; for example, we can now say that behavior is conditionally independent of beliefs given emotions, which implies a different causal chain (Figure 8b) than a model that combines beliefs and emotions additively as in [Oliver 1993].



(a) *Beliefs* and *Emotions* are two factors that (additively) influence *Behavior*

(b) *Emotions* mediate the influence of *Beliefs* on *Behavior*

Figure 8 How do emotions affect behavior?

Since our understanding of the empirically-valid psychological constructs is quite limited at present, Figure 9 shows many constructs such as "goals" and "needs" without any connections to the other constructs. It is clear that "cognitive categories" are central to the model; they trigger the recall of emotions (specifically, somatic markers [Damasio 1994]). Evaluations and attitudes appear not to be direct causes of behaviors, but parallel variables; in some sense they mimic the effect of emotions. However emotions are the primary construct and should be the subject of study; this is contrary to the PAD, DES, and CES models [Meharabian and Russell 1974, Izard 1977, Richins 1997] where emotions are modeled as comprising three dimensions one of which is evaluations.

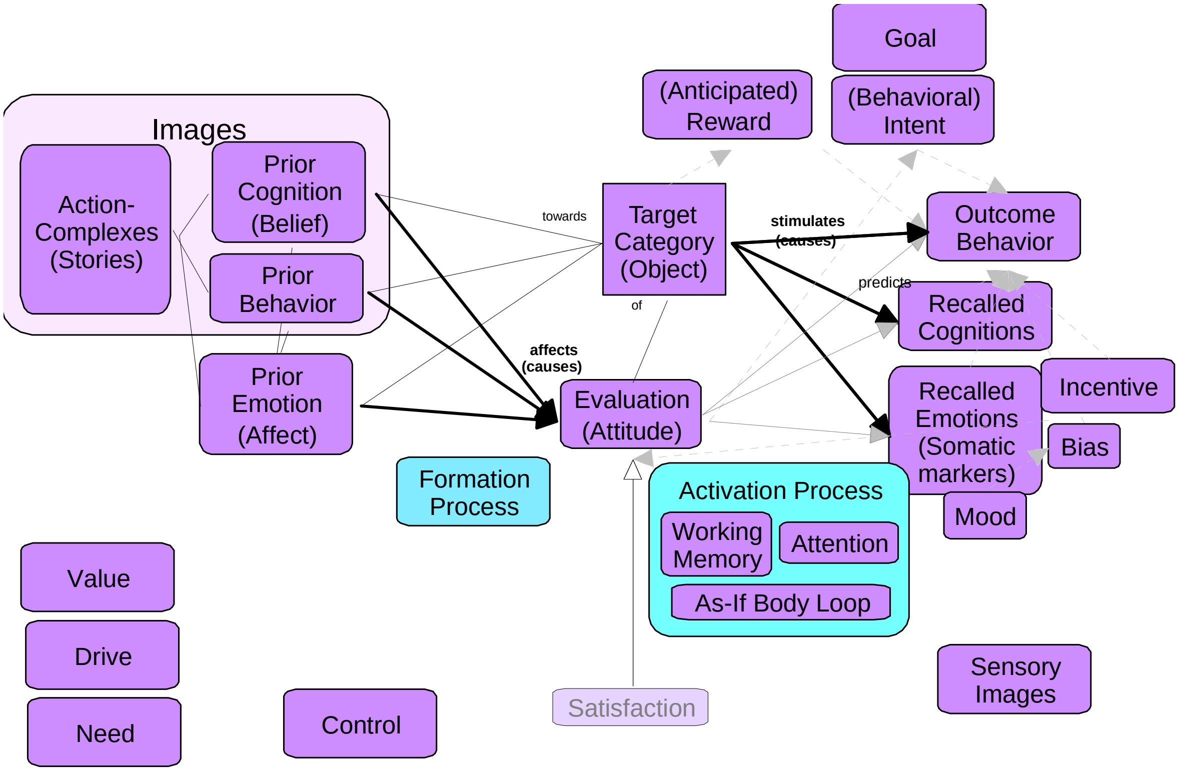


Figure 9 Some important psychological constructs and their relationships

The primary psychological metaclasses are listed below. “Cognitive categories” have not been called out separately since the Objects category previously included in the metamodel deals with them. Although the precise role of “Intent” is unclear---it is well known that intent is a poor predictor of behavior anyway---we have currently retained it on our list.

Emotions

Participant:

- How did you feel then? What did you feel?

Observer:

- What actors/objects/acts/activities/events/states/places/etc., are associated with feelings? (Esp. strong feelings?)
    - What stimulates feelings?
    - What is affected by the feelings?

Beliefs

- What did you think? What went through your mind at that time?
- Special categories of beliefs of interest?
    - Domain independent: Norms / expectations
    - Domain dependent:

Intentions / goals

- How are objects used in achieving goals?
- How do goals involve acts, activities, and feelings?
- Which goals are scheduled for which times?
- What are all the ways goals evoke feelings

3.1.3.3 Special categories

Information and communication

Observer:

- What communication channels and access points exist in the space (e.g., phones, ATMs, catalogs, maps, brochures, guides, signs, advertisements, background announcements or music, concierge, salespeople, Internet etc.)? Where are they located in the space?
- What do they look like?
- What are people reading? Watching? Listening to?
- What did they bring with them? Where did they get it from?
- What did they obtain there? Consume locally? What do they do when they are done? How do they take things home with them? What restrictions are there? Where are people consuming things? And when?
- How are people using communication access points? How often? Who or what initiates the communication?
- How easy are they to use?

Traffic (Motion in Time and Space)

Observer:

- Who/what is moving around in the space?
- Where are the high traffic areas? The low traffic areas? Where are people lingering?
- How is traffic being organized? Is it working?

Tools and technology

Observer:

- What infrastructure has been built into the space (product lookup facilities, notification mechanisms, surveillance)? How does it work? What is embedded? What is visible?
- What is for customers vs. staff?

Food and drink; incentives

Observer:

- What kinds of things are being eaten and drunk? Where are food and drink being consumed? By whom? How are they being paid for? How long are people taking? What else are people doing while eating and drinking? What kinds of meals are they? What hours are meals served? What is being served when you are there? Menus? Service? Advertising?
- What other necessities are supported? What are the incentives provided to participate or linger in this space?

Special psychological/shopping categories

Participant:

- *Topic, interests, goals:*
  - Specific task, specific goals or reasons for task
  - Description of object of purchase interest (or of the class of object if the precise purchase object has not been identified). E.g., "Computers change so fast." "…[frequency of] the latest offering, [is] probably something in the 500s."
  - Expectations about the activity (the purchase-related activity that is the focus of the current episode) and its results (e.g., "If I don't know what computers are available, I expect I'll have to pay more.")
  - Expectations of other participants, and in particular, vendors
- *Outcomes*: results of the episode

Observer, with participant input:

- *Norms of shopping behavior* (Examples of shopping-related behaviors are learning about products, selecting a product, comparing products, etc.): Rules, established processes, what is central, and what is peripheral or digressive.
- *Behavioral styles:* distinctive patterns of micro-level behavior (e.g., concentrated reading of every spec on the label) that may appear often, often in external situations, not just across multiple shopping episodes. (These might be personality traits, not specifically shopping related. The relationship between this and the previous category is that behavioral styles are variants of behavior that are superimposed on the norms.)

- *Norms of interpretation*: norms that shoppers have developed to interpret what happens during shopping episodes. (How do shoppers make sense of events that happen during shopping? What are the cues that people use to determine value?) E.g., “This is much faster than the machine I have at home.”

3.1.3.4 Metarelationships

Metarelationships are relationships between metaclasses, i.e., these are types of relationships that we can always expect to find between elements of data regardless of the subject domain that we study. Having a precise language for these relationships helps us in collecting and organizing many kinds of data.

Generalization: A *Category* X *is a kind of* same *Category* Y.

- Beliefs and Emotions are special kinds of States.
- Behavior is a special kind of Act/Activity/Event.
- An Evaluation is a special kind of Property (or State)

Relationship: A ***Category*** X ***is associated with Category*** Y

Attribution: ***Property*** X is ***an attribute (characteristic)*** of ***Category*** Y

Aggregation: An ***Object / Space*** X ***is in, or is a part of***, another ***Object / Space*** Y.

Sequence: ***Act/Activity/Event*** X ***is a step (stage) in Act/Activit***y Y

Intent: Implicit ***Actor***

Location for action	X is a place for doing Y
Function	An Object/Space X is used for Y
Means-end	X is a way to do Y
Rationale	X is a reason for doing Y

Evaluation: An ***Actor evaluates*** (a special type of ***Act/Activity***) the ***Evaluation*** (merit, worth, importance) of a ***Category*** (typically an Object, but can include abstract entities as well)

Feeling: An ***Actor feels*** an emotion/mood (i.e., is in one of a special set of mental or physical ***States***).

Behavior: An ***Actor does*** an ***Act/Activity***. (I.e., behavior is a special type of act/activity that always involves an Actor)

Belief: An ***Actor believes*** (i.e., is in a special mental ***State*** of holding to be True) a statement (a ***Relationship***, including instances of any Metarelationship, including this one).

Causation: ***Event/Act/Relationship*** X ***is a cause*** of ***Event/State*** Y

Formation: For Actor A, a Belief/Emotion/Prior-Behavior (special kinds of ***Event/State)*** X ***formed*** an ***Evaluation*** V.

Activation: For an Actor A, a stimulus X (a special kind of ***Category***) ***activated*** Belief/Emotion/Behavior (special kinds of ***Event/State***) Y

Correlation: ***Event/Act/State/Evaluation/Relationship*** X ***is correlated with Event/Act/State/Evaluation/Relationship*** Y

Prediction: For Actor A, an ***Evaluation*** V ***predicts*** (is correlated with) a Belief/Emotion/Behavior (***Event/State***) Y.

3.1.3.5 Other comments on the metamodel

The metarelationships described in the previous section often relate more than two metaclasses and the relationships between these elements are often symmetrical. However, many analysis tools such as NVivo require data to be organized hierarchically, i.e., using only a parent-child (containment) relationship to link the data elements being coded. In such situations, we recommend using hierarchy in Figure 10.

Figure 10 Recommended coding hierarchy based on metamodel

- **Actors**
  - Participant
  - Husband / Wife
  - Children
  - Other family (e.g., in-laws)
  - Salesperson, manager, etc.
  - Groups
    - Participant and Spouse
  - States
  - **Psychological constructs**
    - Beliefs
    - Emotions
    - Behaviors (traits)
    - Norms
    - Needs
    - Intentions / goals
- **Objects**
  - Objects
  - States
  - Tools / technology
  - Food / drink
- **Places/Spaces**
  - Traffic
  - Signs
- **Processes**
  - Acts / activities
  - Events, states
  - Times / timelines
- **Overall / miscellaneous**

3.1.4 Sampling decisions

One of the important practical considerations that we needed to address is how much data to acquire and how to ensure that it is representative of the causal mechanisms that are commonly at play in the real world. For example, how many shoppers do we need to interview or track, how should we choose them, and how frequent should we capture data from them?

Three methodological observations help with these decisions. First, we only need to observe the *existence* of the causal mechanisms that might explain the outcome variables; we don't actually need to count and have proportional representation for each of these mechanisms during the discovery phase, since the quantitative phase will do the counting. Discovering a mechanism at play will lead us to add the mechanism to our qualitative causal model; observing it again in a second situation does not add much value (although it would indeed increase our confidence in the mechanism). So we only need the smallest sample size that would reasonably enable us to observe all the important mechanisms at play at least once during our study. While there is no way to guess at the sample size in advance, it is easy enough to estimate it on-the-fly in practice. For example, we may interview one shopper and learn about their shopping experience in detail. Interviewing a second shopper may uncover a different type of shopping experience, and similarly a third. But we will soon find that subsequent interviewees describe experiences that basically resemble those of earlier interviewees. The marginal value of later interviewees in terms of uncovering new causal mechanisms drops sharply, and in our experience we don't need more than 10 participants to uncover the main explanatory variables. That number 10 does assume some degree of underlying homogeneity; e.g., if we studied adults and children, we might need 10 of each. However the numbers are typically in the handfuls or low dozens, not more.[11] An examination of the possible large sources of model heterogeneity in our study suggested the following variables: in-store versus online shopping, gift-buying versus buying for oneself, and high-end designer stores versus low-end chains. In the end we recruited 26 shoppers, which in retrospect provided more than adequate variability for the needs of our Discovery Phase.

Secondly, large amounts of data on some sampling dimensions may be easily obtained with small samples on other dimensions. Even a single shopper whose behavior is tracked for 6 months provides data on a number of store trips. Even though the shopper is the same, the trips to different stores (or even to the same store) are sufficiently varied to yield a lot of useful information on many store-related variables. Thus it is possible to amass extremely rich and large qualitative datasets with only a small sample of shoppers. Our study tracked the 26 shoppers over 8 months and thus acquired such a rich dataset.

Notwithstanding the sufficiency of small numbers, it is still important to know that the small sample is representative of the population. Think for example, that if we recruited 26 TV enthusiasts, we might have ended up confirming our client's theory that TVs are perceived as electronic gadgets, instead of discovering as we did, that TVs are treated as furniture and most shoppers don't care about the fancy features, and couldn't tell a digital

[11] Also see the study by [Griffiin and Hauser 1993] who come to a similar conclusion.

TV from an analog one (Section 3.2]. There are many sampling methods described in the literature, including homogenous sampling, intensity sampling, extreme or deviant case sampling, etc. We chose to use the current gold standard of subject sampling, Random Digit Dialing [USDHUD 2000] to minimize sampling bias. We restricted the sample geographically to a 75-mile radius from our lab so that we could meet them and track them in person rather than by phone.

While it is always important to ensure that the sampling process produces adequate variation in the outcome variable (purchase vs. non-purchase from Retailer A), there is no bias-free way to directly sample that variable itself. Furthermore, any variable that is used to define the sample automatically induces a bias on variables that are causally upstream of the sampling variable. For example, if we only sample (or disproportionately sample) visitors to a particular retailer, we are unable to draw any causal inferences on what drives visits to the retailer. This requires that the sample should be defined via variables as far upstream as possible. In our study, this meant that recruited would be limited to people that were as early in their purchasing cycle as possible, had not yet decided on the particulars of their purchase, had not yet visited any stores, and so on. All the variables that emerged during their shopping experience, e.g., which stores they chose to visit, were uncontrolled and representative of the population.

3.1.5 Observation methods

There is a range of techniques used by ethnographers and other qualitative researchers to acquire raw data during field studies. Focus groups, store exit interviews, mystery shopping, and store walkthroughs are among the most popular. The advantages and disadvantages of these techniques are also fairly well understood. For example, in interviews conducted after a person has finished shopping, their self-reported behavior could be quite different from actual behavior. After all, it is hard for a person to be busy engaged in shopping and simultaneously observe themselves. The same problem afflicts focus groups. Moreover, with focus groups and store exit interviews one cannot observe shoppers in their natural settings, whereas a researcher who shadows a shopper through a store may notice many things that the shopper did not, e.g., the shopper might have missed the products they were looking for due to their placement in the store. Focus groups have the additional disadvantage of group effects; it is hard to avoid biasing one participant with another’s explanations. Mystery shopping and store walkthroughs do provide useful insight, but rely heavily on the researcher’s own theories and perceptions which could be quite unrepresentative of the average shopper. Interviews do provide the shopper’s own perspective, but the flip side is that shoppers’ self-awareness is limited and there are strong elements of rationalization---many people are not comfortable describing the messiness of their actual behaviors and prefer to have their decisions appear to be logical.[12] Moreover, shoppers’ descriptions of their own choices rely on

[12] Indeed one of our 26 recruits absolutely refused to create ThinkAloud recordings after his first attempt at doing so. Listening to his initial recording he was so embarrassed to discover how messy his real thought processes were that he insisted on writing a “diary” as the means of giving us all future data about his experience. The diary turned out to be perfectly neat and logical---and nearly useless for our purposes of watching real shopping behavior.

folk theories of decision making, and are limited by their vocabulary; e.g., consider the limited range of terms available to describe emotions.

Our choice of observation methods was therefore guided by several criteria.

- We wished to gain the shopper's perspective, not the researcher's or retailer's. Therefore mystery shopping and store walkthroughs could not be selected as primary observation methods, although we did find them to be useful adjuncts. These two techniques were particularly useful as `dry runs' at the beginning of the study to get first impressions of how the subsequent participant tracking might evolve.

- We wanted to understand individual behavior, not summaries or generalizations constructed via group discussion. Given individual data, we can do the generalization ourselves, but given data that conflates multiple individuals, it is nearly impossible to work back to individual experiences. Since we could not easily see how to obtain clean descriptions of individual experiences via group settings, focus groups were dropped from consideration.

- We also wanted to know the events that *actually* occurred, separately from the shopper's beliefs about what they *typically* do. For example, in response to an improper interview question about how they typically shop, one of our respondents told us that she usually reads a lot of magazines, checks Consumer Reports, visits many stores to compare products, and so on. However proper interviewing about her past activities subsequently revealed that she had done none of the above for her actual TV purchase.

  This is an important point that illustrates a methodological trap that commonly arises due to differences in research philosophy taken by qualitative researchers from different fields. In anthropology, asking about the "typical" is perfectly fine; indeed Spradley's Grand Tour questions [Spradley 1979] start with them. "Norms" and "culture" are in fact the central focus of study. This frees the researcher to enter the observation scene as a `participant observer', because any perturbations they induce on the behavior of the participants (the shoppers in our study) can be ignored by inquiring about the "typical". In contrast, in most scientific research, any such changes in behavior induced by the presence of the researcher are considered to be a form of "bias", since the aim is to study the actual events that occur on their own without interference.[13] The researcher seeks to minimize the "reactive effect of experimental arrangements" [Campbell et al. 1963] or the "Heisenberg effect" as it is sometimes called, and wants to observe exactly what happened each time. This implies that asking about norms and typical behavior is a poor substitute for actual observation, and would be

[13] Although our study is an observational study, not an experimental design, the issue of introducing bias due to the presence of the observer remains the same. The goal of our Discovery Phase is to provide an accurate accounting of what actually happens in the field. This accounting can be viewed a small-sample-size version of what will be observed again at a much greater sample size during the Quantitative Phase. Obtaining accurate descriptions during discovery is critical to deciding what details to keep and what to leave out when we design our subsequent large-scale data collection efforts.

elicited only if a detailed account of the actual events is impossible to get, or if the purpose of the study is to contrast the respondent's perceptions against actual behavior.

For our study we took the latter approach, working hard to elicit actual events and avoid influencing the respondent with our questions. (Avoid bias turned out to be much harder with the real-time tracking method as we describe below.) We did also elicit information about norms and typical behaviors, in order to experiment with how we might use such information in our models. We concluded that such information was not very useful and was normally discarded, because it helped little with discovering new mechanisms and teasing out cause-and-effect relationships. That learning is itself an important methodological insight. A side note of practical importance is that we found it difficult to re-train ethnographers to avoid the "typically" questions and focus on the actual specifics, because such questions have traditionally been the bread-and-butter of their practice. Although modified data elicitation methods for causal modeling are not difficult to learn, they do require learning because a new way of thinking is involved.

- It is extremely important to know the precise sequence in which events occurred, because that helps disentangle cause and effect. This implies careful interview design to minimize errors during recall, e.g., structuring the interview questions to follow the chronological order of events rather than go back and forth. This criterion also gives us a strong preference for real-time observation methods such as tracking, over retrospective ones such as interviews.

  This criterion was crucial in influencing our ultimate decision to design our discovery phase as a longitudinal study, tracking people over many months of actual shopping, rather than relying on retrospective methods such as interviews.

  This criterion also influenced us to favor the use of the ThinkAloud method as described below.

- It is much more important to obtain a "stream-of-consciousness" dump of raw thought processes than to obtain a structured or logically coherent dataset where the structuring has been done by the participant. Research in psychology tells us that most "thought" is subconscious, and not even linguistic [Lakoff and Johnson 1999]. Conscious thought is only the tip of the iceberg; ideally we need to elicit other kinds of data such as emotions, visual processing, and so on. Given the limited tools available for acquiring such data, especially in naturalistic settings outside a laboratory, our hope was to at least obtain a linguistic dump of raw thoughts that we could subsequently analyze and structure as needed, rather than work with descriptions rationalized by the shoppers. The ThinkAloud protocol [Ericsson and Simon 1993] is an excellent technique for obtaining stream-of-consciousness data. Its main disadvantage is that participants have to be trained to think aloud; without training they tend to "explain" things rather than describe what is happening, or to fall into silences once they get engaged with shopping activities. Also, a few participants can feel self-conscious talking aloud in public, although most quickly get comfortable when we suggest pretending that they are talking on a hands-free cell phone, a phenomenon so ubiquitous today that nobody notices it. The ThinkAloud method has the major advantage that when the participant starts talking aloud, they have no mental bandwidth to explain or rationalize their perceptions; we get good clean data. Because the data is captured in real-time, the recall errors present in retrospective interviews are also avoided, and causality can be teased out more reliably.

- We wanted to ensure *reliability* and *traceability*. After the data was acquired, we wanted multiple analysts to be able to independently look at the same data, in order to could minimize researcher-bias; this is sometimes called inter-rater reliability [Boyatzis 1998]. This required direct capture of raw data, not just reliance on field notes and impressions from the observers. It was also important to us from a scientific perspective to build high-quality scientific models. This meant being able to ground all downstream analysis, both in the Discovery and Quantitative phases on real empirical data, minimizing the layers of "interpretation" that would otherwise creep into the analyses and subtly inject *a priori* theories (see Section 3.1.1 for a discussion of this topic). This in turn required the ability to connect all downstream artifacts, such as the elements of the qualitative and quantitative models, to the upstream raw data that triggered the creation of downstream elements, so that at any time later (when much was forgotten about the field study, and when other researchers join in for the quantitative work), we could always examine any model element or assumption and justify its existence, meaning, relationships, etc., via the observed data. Capturing data in various electronic forms (plus the use of connective tools such as NVivo) was important for achieving such a high level of traceability.

Since we wanted to watch what happened to shoppers over the course of their entire experience, we needed to catch them as early as possible in their purchasing process, and then set up a data capture process by which we could acquire longitudinal data. The recruiting and sampling decisions we made are described in Section 3.1.4. Data capture was designed to utilize a combination of methods, including conversation recordings, ThinkAloud recordings, photographs, weekly debriefs, a post-completion retrospective debrief, and additional structured and semi-structured exercises. The latter were conducted at the end of the tracking period to avoid influencing shopper behavior during the study.

Although we recruited shoppers at the earliest possible stage of their shopping experience, because we had no way to recruit them at the point of earliest awareness of their need for a TV, we still needed to catch up and obtain data on their past experiences. We persistently probed on this during the initial interview ("*Could we go back in time to the earliest moment you can remember thinking about getting a TV? What got you started on looking for a TV?*) and as a result uncovered a lot of causally relevant information going quite far back in time. Indeed, we discovered that the TV purchasing experience begins as long as two years before the "active" shopping phase (Section 3.2.2), not 2—3 months as we had thought.

The raw data was captured using small digital voice recorders with a long recording capacity, which shoppers could stick into a pocket or purse and forget about. The recorders were used to capture both ThinkAloud vocalizations, as well as actual conversations between the shopper and other actors that appeared in various observation scenes, e.g., family members discussing TVs over dinner, conversations with salespeople at stores or over the telephone, and so on. In some cases, our study participants asked permission or mentioned to the other actors that the conversations were being recorded, and in other cases, they did not mention the fact; this depended on the local laws governing the use of recording devices. None of our participants were prevented from using the recorders, at any retailer that they visited.

Video recordings in the field were considered as an additional data capture method, but deemed to be either too difficult or unlikely to be as useful as audio for our purposes. Recording a shopper's experiences on a camcorder requires the presence of a second person to record the video, since it is hard and too disruptive for the shopper to carry a camcorder and shop at the same time. Many shoppers normally shop alone, and it would be difficult for one of the researchers to accompany the shopper without scheduling a trip and thereby changing the shopping process. (Nor did we have enough ethnographers to shop along with all the two dozen participants.) Moreover the forced presence of a second party would probably alter the shopper's behavior, even when the companion belonged to the shopper's family. Miniaturized cameras attached to the shopper's body would be less intrusive but were deemed to be too difficult to conceal and focus on the relevant parts of the scene. In particular, head-mounted cameras (concealed in eye-glasses) appeared to be highly desirable from a research perspective because they capture the focus of attention, but were too large and difficult to use in a natural way. Mounting cameras within the scene itself was another option considered. While it was possible to install cameras at stores belonging to our client sponsor, gaining entry into competitors' stores was a problem. Obtaining data only from our client's stores is relatively useless for predicting the choice of retailer, although the data might be useful for explaining purely within-store outcomes (such explaining whether or not participants talk with salespeople or buy anything on a given trip)---for our client only. Therefore, in-store cameras missed the primary objective of our study. Moreover, we anticipated many serious problems with analysis, such as wading through large amounts of irrelevant footage, or not being able to explain on-camera events because their causes might be off-camera events that might have occurred in the shoppers' lives well before they even visited the store. For all these reasons, installing in-store cameras was deemed to not be worth the effort.

The recruited participants were clearly instructed that they were not required to perform any particular activity for the study, but merely to do whatever they would normally do if they had never spoken to the researchers. In particular they were not required to purchase a TV in order to receive the incentive that they would be given for participating in the study. They merely had to record their thoughts and feelings during and about any shopping related activities that they engaged in. They were also asked to save any artifacts that might help the researchers understand their experiences. Participants were asked to record ThinkAloud conversations while they used computers to shop online, and the URLs visited were logged. The research team subsequently retrieved copies of those Web pages for analysis.

The researchers also collected other artifacts from the participants, such as any notepads on which they jotted down product information, the advertising flyers that they looked at, and so on. These artifacts turned out to be especially insightful during analysis. The participants were also given small disposable cameras, and some participants took pictures of products during their shopping trips, the TV that they purchased in use at home, and so on. These turned out to be somewhat useful. However, more valuable were the field trips that the research team conducted *after* a participant provided data on their experiences, during which the researchers retraced each step taken by the participant to visualize and understand the actual scenes from the participant's perspective.

The incentives for participation were carefully designed to minimize any influence on behavior during shopping for a TV. For example, cash incentives could not be used because participants might view the incentive as being in effect a discount on the price of the TV to be purchased, which might make them buy a more expensive TV than they would have done by themselves. Similarly, any gift which might be viewed as an accessory or complementary product (e.g., electronics) might influence purchase behavior. The incentives finally chosen were to give each participant 3 gift certificates to stores providing unrelated merchandize, e.g., Bed Bath & Beyond, and Home Depot, totaling $350. The first certificate for $100 was given during the initial interview to set up a positive relationship with the respondent, the second certificate after a certain period of participation, and the last one at the end of the final interview.

During tracking, each participant was asked to send in their data recordings every week, by mailing the voice recorder's data chip in prepaid envelopes supplied by the researchers. (Participants were given spare data chips, which were replenished as needed.) Further, a weekly interview was conducted to obtain a "retrospective debrief" of the past week's events; this interview lasted anywhere from a few minutes to an hour depending on the amount of activity that had occurred. At the end of the study, a special 2-hour in person retrospective debrief was conducted during which the participants were asked to recount their entire shopping experience from beginning to end. The weekly and final accounts were contrasted with the real-time recordings whenever possible to compare actual behavior with recalled behavior. This was not always possible, since participants often forgot to turn on their recorders and in such situations only retrospective data was available. When both were available, it was common to find discrepancies, e.g., the participant's recalled explanation of why they ended up focusing on a particular TV (e.g., the prices) was different from what appeared to be relevant factors based on an analysis of the context (e.g., they were admiring the stand that the TV was sitting on, and it was the only TV on such a stand). For the purposes of causal analysis, i.e., for explaining the mechanisms that affect actual purchasing behavior, we relied primarily on the real-time data rather than retrospectives whenever the former was available. The retrospectives, especially, the final debrief, made the purchasing process appear to be a lot more organized than the actual real-time data indicated; therefore the retrospectives were more useful for `summarizing' the purchasing process in participant language, e.g., answer "what does the overall decision process look like?", but not for explanations, e.g., "why did participants choose that particular retailer?"

Our reliance on longitudinal tracking substantially aggravated the risk of researcher-induced bias, because of the need to regularly talk with the shoppers and collect their data. For example, if we ask a shopper whether they had gone online to look at TVs during the previous week, we are in effect reminding them to go online next week; left to their own devices they may not go online for the entire duration of their shopping experience. (This turned out to be the case with many shoppers, as we eventually discovered). So our contacts with the shoppers, and especially the debriefing interviews had to be carefully designed to elicit all the details of their previous experiences, jogging their memory about thing that had already occurred, without triggering changes in (or at least minimizing influence on) their future behavior. Such an interview script, which we call an "ethnographic debrief" turned out to be quite difficult to design, and even harder to execute, especially for traditionally-trained ethnographers used to interviewing respondents about the "typical". Surprisingly enough, the debrief script that we finally devised is extremely simple in appearance, with only three questions used repeatedly (Figure 11); the difficulty in design and execution was in learning what to exclude from the script, and what not to say during the interview.

These "ethnographic debrief" interviews are quite different from another type of semi-structured interview traditionally used in ethnographic research, which we also conducted as a supplementary activity. We called these the "ethnographic life-history interview". This interview comprised the standard grand-tour/mini-tour questions from Spradley [1979] and others, such as "*How would you describe your typical shopping trip?*" "*What role do others play when you go shopping?" "Are there any common emotions or moods that you associate with going shopping?" "Could you describe your best shopping experience?" "Your worst shopping experience?"* And so on. This type of interview was designed to use Spradley's DRS method to uncover the "culture of shopping," participants' perceptions of themselves and their world. The results were highly interesting from an ethnographic perspective. For examples participants viewed their TVs as a comforting companion, who would lull them to sleep in the evenings; a presence in the room, which provided company even when they read a book and didn't actually watch the TV. While this was fascinating, one of our methodological conclusions was that such interviews provided little insight in terms of causal explanations of behavior. For example, a retailer wanting to explain and influence individual purchasing decisions would gain little from such cultural accounts, because the retailer's primary need is for causal variables that they can control and modify to produce a predictable effect on the outcome, which is more easily obtained through detailed retrospective debriefs. Because these life-history interviews focus on the "typical" (and because of the resulting methodological issues that we discussed earlier this section---in particular because we wanted descriptions of actual events more than we wanted perceptions), we conducted these interviews only in the final meeting with each participant at the conclusion of the study.

It is worth noting that observational field trips by researchers (by themselves, not accompanying any shopper) turned out to be of limited value in building causal models. The problem is that watching random shoppers for a few minutes or hours as they walked through a store did not reveal much into the what caused them to visit, what they were looking for, why they were looking for it, what they processed mentally, why they finally bought something or walked out, etc. The behaviors overtly visible in the store provided weak or no clues to what went on invisibly inside shoppers's heads. Given the state of today's mind-reading technology, the ThinkAloud procedure is indispensable in such circumstances when studying other people in the field. Of course the researcher knows what is going on in one's own mind so they can try to analyze their own experiences in the store, e.g., by performing mystery shopping. However during analysis it is extremely

Figure 11 Ethnographic debrief script

1. **Startup**

"During the recruiting process, I understand that you indicated that you are looking to buy a <product of interest>. I was wondering if you could tell me about your experience so far. What shopping related activities have you engaged in so far?"

Check:

- Is the informant talking freely?

"If you don't mind could you elaborate a little on all these things? Could we go back in time to the earliest moment you can remember thinking about a <product of interest>? What got you started on looking for a <product of interest>?"

Check:

- Does the answer reference the <product of interest>? If not, rephrase the original question.
- What events made them think about the <product of interest>? If events are not clear, ask that.
- Is this the very first time they thought about the <product of interest>? If not, ask that.
- Was the act/activity/event was described in sufficient detail? If not, ask "Could you elaborate about what you where thinking back when you got started."

2. **Main questions** (Iterate on the following)

a. (Elaborate on the next thing that happened) Pick the <next event/act/activity> that the informant mentioned. "*You mentioned that <the next thing that the informant said>. I was wondering if you could elaborate on <repeat what the informant said>*?"

Check: Was the act/activity/event was described in sufficient detail? Does the question "How did you feel?" make sense?
If so, pick a <moment in the experience> and ask one of the following questions, as applicable:

b. (Feeling) *"How did you feel when <moment previously mentioned by the informant>*?" or "*How did the <stimulus object previously mentioned by the informant > make you feel?*"

c. (Thinking) Pick an Act, Activity or Event previously mentioned by the informant and ask what the informant thought. "*What went through your mind when <use informant language describing a previously mentioned act, activity, event>?*" *or "What did you think when < use informant language describing a*

hard to separate one's own theories of shopping from one's own experiences, and thus the experience itself adds limited value beyond an armchair analysis that could have been done by the researcher without stepping out into the field. For these reasons, we downgraded the value of independent field trips in building models that involve a significant element of human behavior.

3.1.5.1 Cognitive category maps

Another interview method that we utilized was a structured "taxonomic interview", also conducted at the conclusion of the study. This interview method was designed to elicit shoppers' "mental models" about various things---types of TVs, places to shop, types of shopping activities, and so on. Although the study of taxonomies is common in ethnographic and human-computer interaction research, the interviews we designed were a novel kind, designed to produce "Cognitive Category Maps" as their output. These interviews were designed using insights from second-generation cognitive science, and category theory in particular [Lakoff and Johnson 1999]. Specifically, the notions of "well-defined cognitive category" and "poorly-defined cognitive category" were defined using two major characteristics identified in category theory research: the presence of a unique image, and the presence of a unique set of actions, for a given cognitive category. That led to a set of 5 questions (Figure 12) and an iterative process for quickly eliciting the raw data need to perform cognitive category analysis later and generate a "map" of the shopper's conceptualization of a domain; the analysis process is described elsewhere [Noronha and Kramer 2006]. These interviews were usually `primed' with a checklist of terms (potential cognitive categories) mentioned by that particular participant during previous interviews, to help recall and save time; the interview typically takes less than fifteen minutes.

Cognitive category maps do provide some value towards our end goal of building an explanatory causal model, but are not central for that purpose. Their main value in causal modeling is in identifying the "levels" of each variable. An illustrative example may help with understanding this, as it is related to an important insight from our research. One of the shoppers that we tracked made the following remarks (as part of her ThinkAloud recording) while visiting a store:

> *"Okay. We're at The Wiz right now, which is going out of business and they're having big sales all over the place. We looked at a couple of TVs and we saw the Toshiba. The same Toshiba that I saw at the Best Buy the last time that I went there and it's for like $100 cheaper."*

She therefore thought that the TV was a good deal and went on to buy it. When our research team later looked into the details of her purchase, we discovered that the TV she saw at The Wiz was not the same as the one she had seen at Best Buy, although both were made by Toshiba. Best Buy was selling a less expensive model with fewer features, whereas The Wiz carried the more expensive "Cinema" series from Toshiba. However, our shopper apparently did not notice the difference in models or series, and thought that both carried the "same" TV, a "36-inch Toshiba TV", since the look of the TV, the size, and all other aspects that she noticed were identical. Clearly her mental model was different from that of retailers and manufacturers who are much more sensitive to differences in models, features, etc., and think in terms of SKUs and model numbers when conceptualizing TVs. A consumer who can't tell the difference will think the cheaper one is a better value. A retailer who carries the more expensive model may lose the customer to another retailer who has a cheaper TV with extra features, because the

Figure 12 <u>**Taxonomic interview (for eliciting cognitive categories)**</u>

1. You mentioned <*all the types of TVs mentioned so far, e.g., 27inch, 32inch, HDTV, the TVs with all the bells and whistles*>. What are the other **kinds** of <*TVs*>?
2. What are the **differences** between them?
3. How do these <*TV*s> differ in what you would **do** with them?
4. When you think of <*each type of TV just mentioned*>, does a **picture** pop into your mind? Can you sketch it?
5. **Tabulate** the different types with the differences: How would you characterize <*type of TV mentioned*> with respect to <*differentiating characteristic mentioned*>?
   Checklist: high definition, digital TV, flat screen, projection TV, HDTV, plasma TV

<u>Worksheet:</u>

Type of TV	Size	Mount	Picture in head	Use	…
Flat Screen	Big	Don't think of as mounted	TV I returned. Big and bulky	…	
Flat Panel	Small/tiny	Can mount on wall---can go anywhere	Picture it on the wall		
…					

customer may think that the TVs are identical. Clearly it is important to capture a customer's mental model precisely if we wish to accurately predict their behavior. Since store transaction databases have SKU-level definitions of TVs, attempts to build predictive models of purchase behavior using such data will have to contend with extra noise and modeling error due to their inability to reflect the cognitive categories that actually influence shopper behavior.

This leads us to state the following hypothesis: when the levels of all the variables in a causal model correspond to well-defined cognitive categories, the explanatory power of the model rises substantially. More precisely, if we are trying to predict an outcome Y using an explanatory variable X, and we therefore regress Y on X, if either variable is ordinal or categorical, the variance explained depends on the way the variables are defined in terms of their values. If X is meant to represent the *Type of TV*, defining the values of X by listing all the types of TVs in the retailer's product database will produce a worse model than defining the values of X by listing all the well-defined cognitive categories of TVs. We have not yet experimentally proven this hypothesis; our argument is currently theoretical. It is based on the observation that the first way of defining *X* introduces unnecessary noise (i.e., distinctions between TVs that don't make sense to consumers, and therefore do not influence their behavior) into the model.

The main value of cognitive category maps lies outside causal modeling, and is not directly relevant to the focus of this paper. However, their value is worth noting because most situations that require business insight typically do benefit from building these maps as a side exercise while working towards the causal model. The side benefits of having a material representation of what is in the customer's head are numerous. Most important is the design of marketing communications---one can now speak the language of the customer, by limiting one's language to well-defined cognitive categories, and this is valuable input to the design of advertisements and Websites. This also has immediate implications for optimal design of product lines and product catalogs [Noronha and Kramer 2006]. The maps also have implications for up-sell and cross-sell: salespeople are likely to fail at up-selling if they attempt to cross cognitive category boundaries (e.g., if they try to sell a projection TV to someone looking for a mount-on-the-wall TVs).

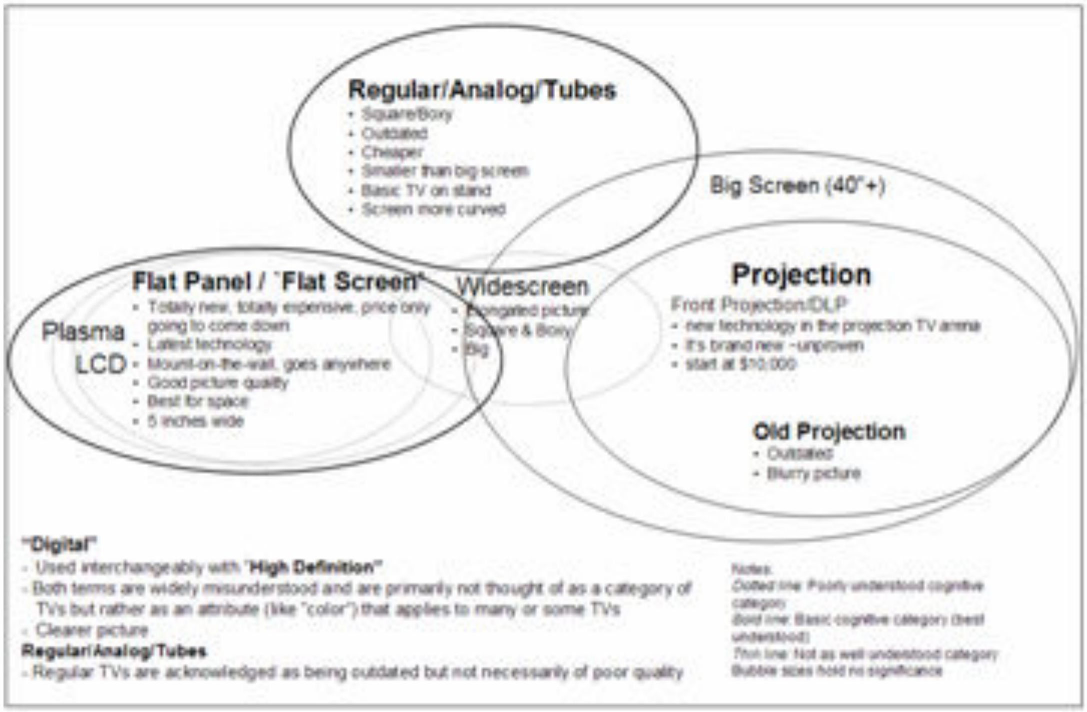

Regular/Analog/Tubes
• Square/Boxy
• Outdated
• Cheaper
• Smaller than big screen
• Basic TV on stand
• Screen more curved
Big Screen (40"+)
Flat Panel / "Flat Screen"
• Totally new, totally expensive, price only going to come down
• Latest technology
• Mount-on-the-wall, goes anywhere
• Good picture quality
• Best for space
• 5 inches wide
Plasma
LCD
Widescreen
• Elongated picture
• Square & Boxy
• Big
Projection
Front Projection/DLP
• new technology in the projection TV arena
• it's brand new –unproven
• start at $10,000
Old Projection
• Outdated
• Blurry picture
"Digital"
- Used interchangeably with "High Definition"
- Both terms are widely misunderstood and are primarily not thought of as a category of TVs but rather as an attribute (like "color") that applies to many or some TVs
- Clearer picture
Regular/Analog/Tubes
- Regular TVs are acknowledged as being outdated but not necessarily of poor quality
Notes:
Dotted line: Poorly understood cognitive category
Bold line: Basic cognitive category (best understood)
Thin line: Not as well understood category
Bubble sizes hold no significance

Figure 13 Cognitive category map: "Types of TVs"

3.1.5.2 Alternative ways of designing the discovery phase

Despite the value we obtained from longitudinal tracking, for pragmatic reasons we cannot recommend it as a necessary component of our Discovery Phase methodology in all business situations. Given the speed at which businesses work, there is a need to construct causal models quickly and deliver results within a few months, if not weeks. Longitudinal tracking takes as long as the time needed for a small sample for actors to complete their decision making process. If the purchase cycle for TVs is on average a year, the tracking must last several months so that at least a few participants complete the process. This length of time, just for the data collection part of the Discovery phase, never mind analysis and Quantitative Modeling, is often unaffordable unless the client really takes a long-term view. This is not a problem when decision processes are short, e.g., purchasing of clothes or small kitchen appliances. It is easy to use longitudinal tracking in such situations.

When the data collection period must be shortened, there are several alternatives to longitudinal tracking that can be used, to achieve the best tradeoff between data quality and speed. In our experience the `ethnographic debrief' interviews are the best choice in most circumstances. Lasting 2 to 3 hours per participant, dozens of participants can be interviewed (depending on the number of researchers conducting the study) and data collection completed within a week. Each interview elicits a complete description of the participant's experiences and decision processes, minimizing recall error to the extent possible. The price of course is that analysis becomes less reliable since the sequence of recalled events has to be carefully disentangled in order to tease out cause and effect mechanisms. However, occasional errors in analysis might be considered an acceptable tradeoff for the fast delivery time.

3.1.6 Qualitative analysis

To complete the Discovery Phase, the raw data obtained via field studies must be transformed into insightful knowledge. Such knowledge may be of two types: (1) Insightful observations or conclusions that can be immediately interpreted and used to draw business conclusions, e.g., most shoppers do not understand the differences between "digital TVs" and "regular TVs", and therefore advertising campaigns around the word "digital" are likely to be ineffective. (2) Observations that suggest insightful theories or hypotheses, that should be further analyzed in order to systematically draw business conclusions. For example, a visit to a Best Buy store appears to trigger a visit to a nearby Circuit City store, and therefore Visits to Competitors may be an important driver of Visits to our Client's store. The importance of this new mechanism can only be measured by a large-scale statistical study; the main function of the Discovery Phase is to discover the existence of such mechanisms.

While both type of qualitative knowledge are important, the second type is of special interest because it enables the building of causal models that ultimately answer the central business question: what are the critical factors that drive purchases? The challenge in designing a qualitative analysis methodology is to ensure that there is a rigorously scientific transition from raw qualitative data to structured inferences that are suitable for input into statistical models, and in particular, to causal models.

The fact that a subsequent Quantitative Phase will be the `consumer' of the analytical results from the Discovery Phase induces specific requirements on the nature and format of qualitative analysis. What kind of qualitative input do causal models need? There are at least three types of information that must be produced:

(a) Causal models require help in the choosing the variables to be modeled. Not only must the Discovery Phase help ensure that all the relevant variables are present, but it should also help in specifying the possible values that the variables might take, i.e., the granularity of the variables (see Section 3.1.5.1).

For example, consider the variable "Price". While at first glance, this sounds like a straightforward construct, it has in fact many interpretations and formulations, and in the end several distinct variables must be included in the model. Apart from the obvious "Price of the final TV under consideration at retailer A" (which itself has multiple instances corresponding to the multiple retailers), there is the "Price range" or budget of the shopper, which limits the set of TVs that the shopper even considers seriously; the "Price sensitivity" which represents how much the shopper cares about price relative to other attributes of the purchase such as picture quality; the "Price threshold" at which the shopper does not really perceive a difference in price (e.g., if the shopper considers a $429 TV no different from a $449 TV, the $20 price difference is below their threshold), and so on. Each variable is formulated and measured differently from the others, and plays a different role in the model.

(b) Causal models require guidance in terms of the mechanisms that connect the variables to each other.

In the above example, the *Price* of the final TV might affect the *Liking* for the TV, or the decision to *Purchase*. However the *Price range* has a different effect: it limits the *number of TVs considered*, but does not affect *Liking* and will not affect *Purchase* unless there are no TVs available at all within the price range. The *Price sensitivity* moderates the effect of *Price* on *Liking* and *Purchase*. The *Price threshold* only affects the *Purchase* when two TVs are being *compared* from different retailers.

Thus the qualitative analysis must identify all the first-class causal mechanisms (see the discussion at the beginning of Section 3.1) that may ultimately influence sales. Not only must the qualitative analysis detect and call out mechanisms, it must also provide guidance on the reliability and validity of these mechanisms. For example, some mechanisms may be observed directly via a field study, and others may be hypothesized in the literature. When the Quantitative Phase needs to make modeling assumptions, the qualitative analysis results should prioritize the assumptions in sequence of trustworthiness. (See the next section on ways to document the plausibility of causal links, and Section 4.1.2.3 for a system for injecting assumptions).

(c) Causal models also require guidance in terms of the functional structure of the mechanisms, since a lot of relationships are nonlinear.

For example, does a higher *Price* proportionately reduce the likelihood of *Purchase*, or does the shopper's price sensitivity to price change at different points on the price scale? If we increase the *Frequency of advertising flyers*, will it proportionately increase the probability of *Visit*, or is there a saturation effect? If there is a saturation effect, what is the shape of the relationship between *Frequency* and *Visit*? Qualitative analysis of field data or the literature may suggest using a logarithmic function, following Weber's law. Similarly, qualitative analysis may suggest using a multiplicative (conjunctive) form to explain the effect of screening attributes (Section 4.4.3.1) on *Serious consideration of TVs*.

Thus, the central function of our qualitative analysis methodology is to uncover the above types of knowledge from field data and enable the modeling of cause and effect. The primary data structure that we produce is the Qualitative Causal Model. As we mentioned when we introduced these models in Section 2, it is important to define these models using the semantics of Pearl [2000] and not confuse them with the influence diagrams and causal networks described in the traditional qualitative analysis literature (e.g., [Miles and Huberman 1994]), because the former has rigorous quantitative properties as well as qualitative meaning.

The following section will briefly describe practical methods for building qualitative causal models. The subsequent section will then return to the more general question of role of traditional qualitative analysis methods such as Spradley's DRS, domain analysis, etc.

3.1.6.1 Building Qualitative Causal Models

In designing this analysis process, we have experimented with many techniques for building qualitative causal models, utilizing taxonomic and domain analyses, process models, generation of causal model fragments followed by assembly of fragments, individual case analysis followed by generalization, simultaneous multiple case analysis, and so on. We found that most of these methods create a large analytic workload but are only partly useful for building causal models. With a large amount of field data to analyze, and given the fact that data analysis typically consumes 5-10 times the effort required to acquire the data, it is important to strip down traditional analysis methods to the minimum required to extract the needed qualitative causal information. This issue is less of a concern when the analyses are used for other purposes as well, e.g., building mental models of consumers. However, in most business situations there is considerable pressure to quickly and efficiently produce the information needed for the primary end-goal (determining what drives sales), without sacrificing scientific rigor. The overall methods described in the traditional literature (e.g., the inductive/deductive analyses described by [Miles and Huberman 1994, p. 155] for causal networks) are ill-suited for this purpose, although some of the specific heuristics (e.g., [Miles and Huberman 1994, p. 146]) are very useful.

The main difficulty arises from the intrinsically `backward' nature of causal analysis: one needs to work back from the primary outcome under study (e.g., the purchase decision) to its ancestral causes. A forward or fragmented approach will result in the analyst wasting a lot of time on model fragments that do not ultimately have a lot of influence on the primary outcome. However, reading a raw transcript backwards (e.g., from the paragraph at the end where the purchase finally occurs, backwards to the beginning of the transcript where the shopper describes how they initially got interested in looking for the product) is mentally difficult. This is especially hard for qualitative researchers who have not been trained in quantitative methods as well, because the notions of a "variable" and a "mechanism" which are required for formulating model elements add complexity to the existing effort of interpreting the transcript.

In our experience, the best methodological compromise is to utilize a two-step process. First, build a descriptive "process model" which simply lays out the flow of events and activities. This is best done as visual diagram showing who did what and when, and what other events occurred in the process. Analysis is a `forward' process, requiring a relatively easy reading of the transcript and summarization. Graphical tools such as NVivo substantially reduce the effort involved in coding and diagramming. The semantics of a process model are essentially the same as that of activity diagrams in the UML language [Booch et al., 1999]. However, the ethnographic principle of solely utilizing participant language to define the model elements stays paramount. The granularity of description must be both fairly detailed and coarse at the same time: the process model must identify the major `phases' of the narrative, in order to help understand the `big picture' or the `overall decision process' (e.g., "trip to Best Buy", "checking the ads") and at the same time provide enough substance to support the transformation to causal model in the next step (e.g., "Here's a Sharp…pretty nice…flat screen---its $899"). It is possible to start with an initially coarse model and gradually add detail as needed; tools like NVivo become indispensable for this, because they enable the analyst to jump from a node in the diagram to the grounding text in the original transcript. Process models are clearly useful by themselves because of their descriptive power. "What happened" is conceptually a more basic prelude to "why did it happen".

The second step is to morph the process diagram into a qualitative causal model. The reason why it is useful to create a process model first is that causality must respect time. In other words, although temporal sequence does not imply causality, any causal hypotheses we make must satisfy the empirically observed temporal sequence. Therefore, the left-to-right temporal ordering of variables drawn in a process model must be consistent with the left-to-right cause-and-effect chaining in a causal model. The process of creating a causal model is thus simplified. In essence, all the links in the process model have to be examined as to whether they imply causality, not just temporal sequence, and must be adjusted accordingly. Nodes that turn out to be incidental (i.e., occur in the temporal narrative, but do not appear to have a causal influence on the outcome) will be dropped. Nodes may be added as a result of a more detailed examination of the transcript text associated with each variable, to find additional causes that might have been missed in the first round. Analysis for this step is a `backward' process starting from the primary outcome and working back to the ancestral causes. However, this is mentally much easier to do because we already have a full (process) diagram laid out, the big picture is clear, and we can work back piecemeal. The analysis is now `local' to each variable in the model, and it is not necessary to keep the entire narrative in mind when analyzing the causes of a given variable.

Thus the two-step process substantially reduces the cognitive load on the analyst, although the process model itself may not be of primary importance. In practice, we've found the process model quite useful in presenting results to our client, because people psychologically need narrative descriptions of the world, regardless of whether any useful business implications can be drawn from the narrative.

The two-step process, while reducing analysis effort, also has a methodological implication with respect to generalization. It is hard to simultaneously keep in mind the narratives of multiple shoppers and try to figure out causal structure. Therefore qualitative causal models are built one case at a time. Placing several of these cases on a wallboard helps visually identify structure common to all the cases. Generalization is then done by starting with a typical individual's model and adding in variables and links from the other models. This process works relatively easily because the functional interpretation at each node is by default disjunctive: variable Y is caused by X_1 or X_2. E.g., a shopper may have *Visited* a store because of an *Advertisement* or because of the store's *Location*. Finding more variables that might have caused a visit just implies that more X variables are added to the model and linked to the Y.[14] Thus it is possible to just copy the new variables and links from the other models into a generalized model. In more complex situations, although the variables are easily copied over, the links must be carefully checked to see how their behavior interacts with other preexisting links. In our experience there is no disadvantage to performing generalization after modeling individual cases, apart from some possible redundancy when two respondents turn to have similar behaviors.

In terms of content and format, the model usually documents only "first class" causal relationships, i.e., relationships between variables for which the causal mechanism was clearly identified during the qualitative fieldwork (plus a few exceptions where the hypotheses came from the Client). It does *not* specify all the *plausible* causal relationships, i.e., relationships where we can easily suggest a hypothetical mechanism that may exist in the real world, but we did not actually observe during the qualitative research, and are therefore in a class that is much less trustworthy. The value of this approach is it provides a rigorous basis for introducing assumptions during quantitative causal modeling in a disciplined manner: initially introducing only the most defensible assumptions, and slowly introducing others, stopping when new assumptions are no longer needed to obtain a complete model. Thus the strength or defensibility of the quantitative model is maximized, making effective use of a huge amount of qualitative knowledge. First-class or `definite' causal links are rendered as thick blue arrows. When second-class assumptions are used, they are drawn using thin blue arrows. Grey bidirectional arrows indicated common-causes. A handy tip is to use the "layer" mechanism of modeling tools such as Visio to hide all but the definite causes when clarity of visualization is needed with a large model. Another tip is to annotate the "properties" of a link with the actual verbiage used by participants, as a way of documenting the causal mechanism and increasing the transparency of the model.

3.1.6.2 Other qualitative analyses

Given the central role we have given to qualitative causal analysis, what is the role of other qualitative analysis methods such as thematic analysis, taxonomic analysis, and so on, which form the backbone of research methods in many of the social sciences?

[14] Suppose X_1 -> Y for individual 1, and X_2 -> Y for individual 2. Although X_2 has not been identified as a cause for Y for individual 1, it is very likely to be a possible cause that was just not empirically observed for the particular individual. Mathematically, X_2 is hidden within the error term u in $X_1 + u$ -> Y which represents the set of all un-modeled causes. Therefore changing the model for individual 1 to $X_1 + X_2$ -> Y is semantically consistent, and corresponds to an interpretation in which X_2 has been separated out from u.

Our experiments with thematic analysis [Boyatzis 1998], memoing [Emerson et al. 1995] and related methods proved to be significantly less fruitful. The primary issue is the lack of scientific rigor in how themes are developed. We observed some of our analysts who were interested in the subject of human communication readily wrote memos on the conversational patterns between shopper and salesperson. Others interested in the concept of cultural identity readily generated theories on `shopper identity'. We were left with the feeling that analysts put more of their own thinking into the memos than they actually extracted from the real-world data. There was little evidence that any of this mattered much for the focal question, explaining the purchase outcome. It turned out to be more fruitful to direct analysts' attention to the causal mechanisms that directly link the observations. Our conclusion is not that thematic analyses are entirely useless, but that their open-ended nature predisposes the analyst to wander off in a personally preferred direction (which is not a disadvantage if the purpose is undirected academic exploration, but is a problem if there is an end-goal to the study, such as thoroughly explaining and influencing an observable phenomenon such as sales), and the lack of methodological specificity in how themes are identified prevents the replicability that is needed for rigorous scientific inference. Even supporting tools such as our metamodel do not significantly mitigate these deficiencies.

Spradley's developmental research sequence methodology [Spradley 1979, 1980] is much more useful, especially his domain analysis, taxonomic analysis, and componential analysis. Our metamodel (Section 3.1.3), along with tools such as NVivo, greatly help and enhance these analysis methods. We found domain analysis useful for getting a deeper understanding of the products themselves, the settings in which the products are used, role played by these products in the consumer's life, perceptions about retailers, and numerous other things germane to the subject of television purchases. We enhanced taxonomic and componential analysis utilizing concepts from cognitive science (Section 3.1.5.1), which greatly increased its value and range of business applications (e.g., to marketing communications, sales strategies, promotions, and so on). All of this produced substantial insight for the researcher and produced material of the quality suitable for writing an ethnography of shopping. Despite the intellectual gains however, it is hard to systematically convert this insight into business inferences and recommendations with any semblance of scientific rigor. In a commercial business setting with substantial pressure to produce a lot of insight with the least possible effort, it can be hard to justify the many person-months of effort involved. As a contrast in efficiency, consider the amount of insight and knowledge transfer that results from taking Client executives on shop-along field trips.

3.1.7 Adapting the discovery phase methodology: some design tradeoffs

In today's fast-moving business environments there is a need to compress the time required to execute any research methodology to its absolute minimum. In fact many strategy consultants expect their engagements to last no longer than 6-12 weeks, or at most a few months, in order to meet client pressures. In such environments, it is difficult to fit real-time ethnographic techniques into a qualitative research phase that can be at most 2 months long. Some tradeoff must be made between the execution time of a technique, and the quality of data it produces.

Real-time tracking is the most powerful technique for eliciting high-quality causal information because the real-time character of the data provides reliable temporal sequence information. Execution time with this depends on subject matter under study. For products or services whose entire purchasing cycle spans a few weeks, the entire experience can be tracked, but for the multi-year time span of TV shopping, only a small portion of the experience can be tracked. Of course it is possible to obtain data on the different portions of the shopping experience by recruiting shoppers at different stages of their experience---some shoppers can be tracked during their initial shopping phase, and others during their final purchasing activities. However it is hard to piece together data from different shoppers, and especially hard to draw causal inferences that span the different phases. Thus real-time tracking works best when the entire experience can be tracked for each unit of analysis.

As an alternative, or as a complementary technique, we found retrospective "ethnographic debrief" interviews the next best choice. For example, we tracked many shoppers during their active "investigation" phase in real-time, and retrospectively debriefed them about their earlier phases. These interviews take only a few hours per respondent and when executed correctly provide fairly high quality data within a week or two. Relative to real-time data, there is considerably greater omission of detail, confusion of temporal sequence due to recall errors, rationalization, linguistic limitations, and other deficiencies. Despite its weaknesses, we recommend this as the technique of choice when real-time tracking is impractical.

In our experience, third party observations (e.g., researchers making field trips to stores and recording their own observations), focus groups, and similar techniques provide significantly lower quality data for the purposes of causal modeling, and do not offer sufficient time savings relative to ethnographic debriefs to be worth the trade-offs. Thus we do not advocate those techniques as alternatives.

It is worth considering emerging computational techniques to automatically acquire and analyze rich data such as video recordings, a concept we have dubbed the 'virtual ethnographer'. Currently the techniques are not very powerful in terms of automated analysis; the features that are extracted from the data provide little causal information. Manual analysis is extremely expensive---it typically consumes 5—10 times the length of the data recording depending on the medium, with videos are at the high end. Therefore, as of this writing we do not have any practical alternatives of this type to recommend, although we expect useful applications to emerge, e.g., for cognitive category mapping, over the next few years. Also, as qualitative analysis tools such as NVivo become more robust, it becomes possible to maintain greater traceability from raw data to inferences (e.g., from text transcripts or video segments to the corresponding part of the qualitative causal model). Such an improvement in traceability will substantially improve the rigor and quality of the methodology, something that is usually sacrificed today under the time pressure of business environments.

3.2 Findings from the Discovery Phase

From a theoretical perspective, perhaps the most important finding from our Discovery Phase research is rather abstract, viz., that quasi-deterministic patterns of cause and effect govern many aspects of empirical phenomena such as human decision making, and capturing these patterns greatly helps us model and business outcomes as sales. We will describe these patterns in Section 3.2.3, and develop their theory in Section 4.1.4.

From a more practical perspective, there were two types of insights gained. The primary product of the Discovery Phase was of course the Qualitative Causal Model of Consumer Purchasing Behavior. Because of its size, the model is not reproduced here; see [Kramer & Noronha, 2003]. The model encoded numerous variables that potentially influence the purchasing decision, the mechanisms that connect the variables, and hints about the functional forms that may be needed during the Quantitative Phase. Even in purely qualitative form, the model demonstrated its value to our retail client by allowing them to test their strategies. For example, our client asked us "What is the impact of TV brand on the purchase decision? In particular, what would be the effect on sales of running a "Sony Month" promotion?" By tracing through the causal model, we were able to demonstrate the mechanisms involving TV brand were too far removed to have any significant effect on sales, dissuading our client from running the promotion. This qualitative inference was later confirmed using quantitative data. Another example was a question from our client whether a larger store format with many more products would be successful. Their theory was that the added convenience of "one-stop shopping" would boost visits, and therefore sales. Tracing causal mechanisms in our model suggested that any convenience effect was likely to be slim.[15] None-the-less, our client went ahead with the new store format, but as of this writing publicly available reports do not indicate a significant rise in sales. These examples illustrate the value of qualitative causal models for strategic analysis and scenario analysis.

3.2.1 TVs are not male-oriented "gadgets"

The second type of insights gained during the Discovery Phase comprises observations that contradicted our client's prior beliefs about TV purchasing. For example, TVs were (and even today are) widely perceived in the retail industry to be a male-dominated purchase, driven by a desire for "gadgets". This belief is a natural consequence of retailers' classification of TVs as "electronics" or "gadgets", and gadgets are thought to be a male preoccupation. Much retailer advertising is designed according to such assumptions, e.g., the flyer in Figure 14.

Our domain and taxonomic analysis revealed that these beliefs were in error. TVs are "furniture", not gadgets. Indeed TV purchasing is often triggered by a home remodeling activity such as redecorating the living room, converting a basement to a den, or purchasing other furniture such as `entertainment centers'. Homeowners are quite concerned about the "look" of the TV in their entertainment center or how it matches their living room furniture. Women participate as much as men in shopping for TVs. Indeed, couples often shop together, or consult each other before the final purchase decision. The implication for retailers is that TVs should not be marketed in ways similar to gadgets such as DVD players or IPods; rather, TVs should be marketed in ways similar to home appliances. Instead of a sports screenfill, content addressed to women should also be used. Another suggestion is that TVs should come with replaceable facades in

[15] Since our model was for TV purchasing, but the question was about a much larger range of products, we were making a leap of extrapolation; we were thus less confident about our inference.

much the same way that kitchen appliances come with replaceable front panels which can be matched to the room's décor. TV advertising should present TVs in different settings and room arrangements. There are cross-sell opportunities to be exploited between TV and furniture sales. Clearly, such analysis has fundamental implications for retailers' TV marketing strategies.

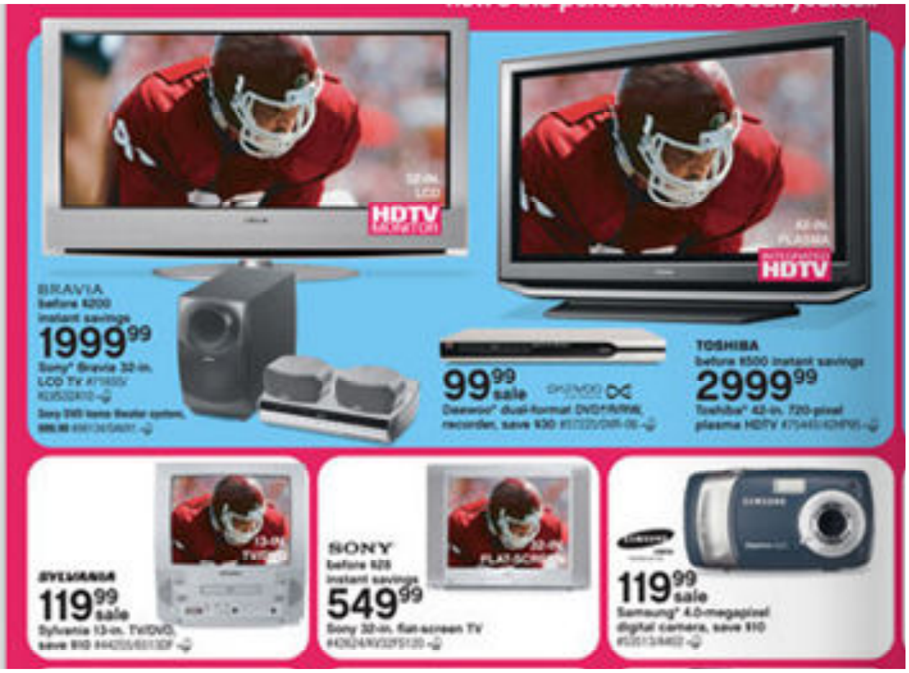

HDTV MONITOR
BRAVIA
1999 99
Sony® Bravia 32-in. LCD TV
HDTV
42-in. PLASMA
99 99 sale
Daewoo® dual-format DVD±R/RW, recorder, save $30
TOSHIBA
before $500 instant savings
2999 99
Toshiba® 42-in. 720-pixel plasma HDTV
SYLVANIA
119 99 sale
Sylvania 13-in. TV/DVD, save $10
SONY
instant savings
549 99
Sony 32-in. flat-screen TV
119 99 sale
Samsung® 4.0-megapixel digital camera, save $10

Figure 14. TVs are (incorrectly) marketed as gadgets targeted to males

3.2.2 The Consumer Decision Process

Another set of valuable insights pertain to the marketing notion of a consumer "decision process" (or "shopping experience"). There isn't one. There are indeed many events and activities that lead up to (and sometimes causally influence) the final event of taking a TV home, but it is difficult to frame the collection of events as a single "decision process" or "experience". Since multiple shoppers participate in and causally affect the final choice, is the child who rejects a particular TV a part of "the" decision process of the parent? If we expand the definition of "the decision process" to include the processes of multiple actors, do we include the retailer who chose to put that particular TV on display in the store? How about fellow shoppers in the store, or on Web sites like BizRate.com whose opinions influence each other? In defining the scope of "the consumer decision process" it does not seem appropriate to either focus on the perspective of a single shopper, or to include every agent who causally influences the decision. We did not even find a "primary decision maker" since co-shoppers often participated equally. This is purely a conceptual difficulty arising from a poor choice of constructs; there no empirically well-defined construct called "the shopping experience" that influences purchase outcomes, nor is there a single "decision process" in which a shopper "becomes aware of a need, finds alternative choices, compares alternatives, and selects an alternative" as marketers tend to theorize. These oversimplifications lose information that is critical for causal modeling. For example, in assuming that shoppers "compare alternatives" there is an implicit assumption that the shopper considered all the TVs in the store. Instead we have observed shoppers missing an entire aisle full of TVs simply because the shoppers did not notice that there were more TVs round the corner. While a "choice" is implicitly made in the mathematical sense that a specific TV (from another aisle) ultimately gets selected from the many available in the store, there was no deliberative "choice" or "decision" to exclude the TVs that were missed. Capturing these behaviors is important for improving the model's explanatory power.

The above conceptual difficulties imply that precise models of consumer behavior will not directly represent the influence of constructs such as "customer experience" on sales, and may not describe "the" decision process. This is not an impediment to building a causal model of sales because all influencing actors and variables are valid inputs to an empirical model. Thus, the above difficulties are methodologically irrelevant from a causal modeling perspective. The difficulties are only important to the extent that today's marketers utilize popular constructs such as "experience", "process", "choice" and so on, in the construction of their strategies. Such constructs are crude approximations, and should be replaced by the empirically-accurate variables that turn out to have the greatest causal influence on the purchase outcome.

Even though the notion of an 'experience' or a `decision process' is not directly useful for causal modeling, marketers have a psychological need for a simplified account of all the relevant events and activities that ultimately affect the purchase decision. An ethnographic organization of these events and activities can provide insight beyond the generic marketing theories described above. By "ethnographic", we mean here that the entire set of events and activities is structured solely via participant-language, in the spirit of Spradley---not using marketer-language such as "wants and needs", "alternatives", "choice", "decision", etc. Such a "customer's-eye" view of the TV purchasing process is rendered in Figure 15.

As the figure illustrates, the portion of the shopping experience that marketer's typically refer to as "the purchase process" roughly corresponds only to the third major phase, viz., "Investigating". Prior to this, there are two other major phases of shopping, "Initiation" and "Getting ideas". Indeed these two phases consume the vast majority of the time involved in shopping for TVs, to the surprise of our client. Our client had estimated the "purchasing lifecycle" for TVs to be about 2 to 3 months in duration, and one of our findings was that the duration is closer to 2 years from the initial awareness of need. The duration is indeed shorter (on the order of days or weeks) when an existing TV in the house suddenly breaks down and needs to be replaced. However it is more typical for an initial interest in purchasing a TV to be triggered by some such event such as seeing a neighbor's new TV, and then be followed by a long period of passive "getting ideas" about what kind of TV one would like to buy, and persuading one's spouse that the purchase is necessary ("lobbying"). Other projects and priorities prevent `active' shopping until some event finally triggers entry into the third phase. Indeed, remodeling projects and furniture purchases, which eventually result in the purchase of new TV, can act as purchase-inhibitors during the first two phases, since the size of the TV or the looks of the TV depend on other remodeling choices, and the TV can be purchased only when the painting has been completed and the furniture is ready. The `passive shopping' activity in these two phases contains many events that trigger, inhibit or influence the events in the final active-shopping phase, and retailers' strategies must find ways to influence those early events.

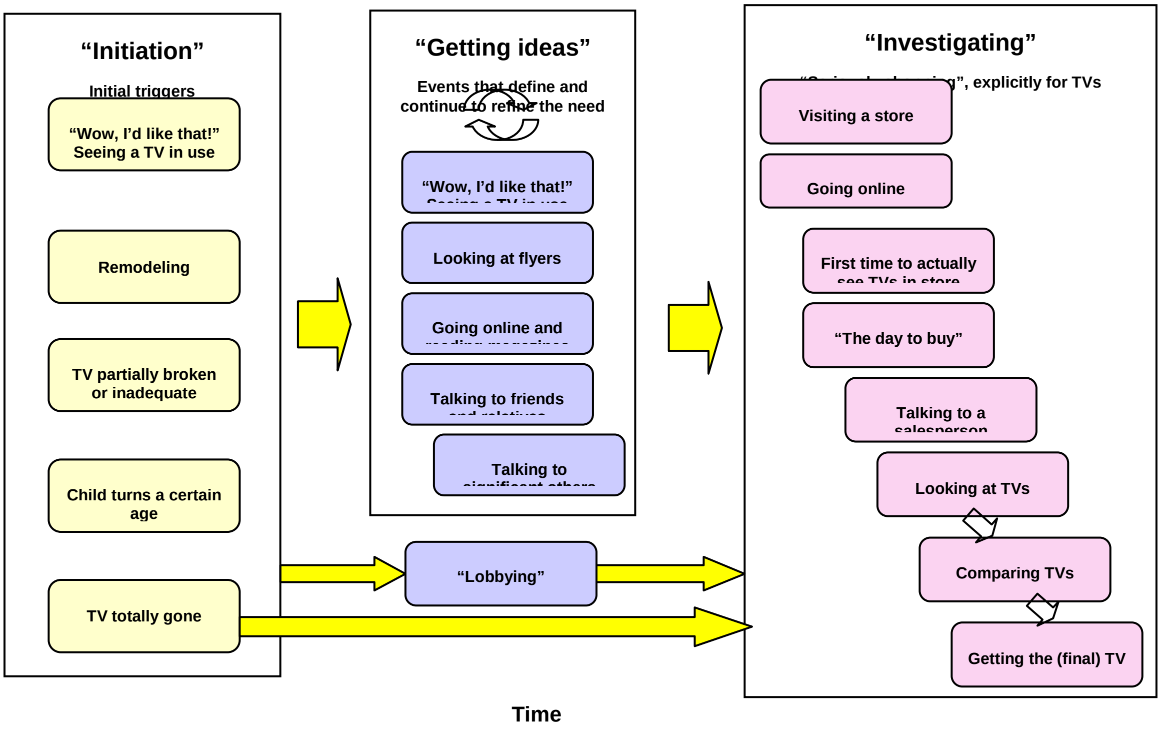


Figure 15 An ethnographic breakdown of the shopping experience

It is worth noting that the long duration of the entire purchase process was uncovered only through qualitative research. In the subsequent quantitative survey, shoppers reported short shopping experiences (1 day to 3 months). The discrepancy is in part due to the broken nature of the "experience" construct---it is hard to define the beginning of the experience, and people often measure it from the time they first visited a store explicitly for TVs, which is a final-phase event. Only through thorough interviewing did we uncover the earlier two phases, which illustrates how survey questions asking respondents to make complex judgments (in this case time estimates of their entire shopping experience) can be relatively unreliable compared to qualitative research.

3.2.3 Quasi-Deterministic Models

One of the most interesting and potentially valuable discoveries that we made during this study is a particular pattern of relationships between model elements that we have termed `quasi-deterministic' relationships, for reasons that will be explained shortly. The significance of this pattern is perhaps best understood by analogy with another functional pattern that is well known, namely the logit transformation. Researchers familiar with choice modeling are well aware of Dan McFadden's Nobel-prize winning work and the importance of logistic models: these models enable much better description of choice behavior than linear models, because the `S' shape of the logit transform more truly reflects the nonlinear relationship between a discrete outcome and continuous independent variables. Furthermore, logit/probit models have a good theoretical interpretation using the notion of a propensity-to-choose latent variable. These constructs have now become the standard tool for building discrete choice models [Ben-Akiva, McFadden, et. al., 2002]. Just like the logit function, the notion of a quasi-deterministic function turned out to provide a much better means for modeling nonlinear choice behaviors. We provide a qualitative introduction in this section; a more mathematical formalization appears in Section 4.1.4.

During our qualitative research, we took a theory-less approach, being completely agnostic about the functional relationships between variables that were being discovered as potentially relevant for inclusion in the model. In particular, we avoided *a priori* assumptions about linear, logistic, or any other functional relationships between any pair of variables, preferring to let the qualitative research reveal any hints about the form of the relationships that would be required later for the quantitative phase of our methodology. We realized that this could potentially create problems during quantitative modeling, since the variables so chosen could turn out to have quantitative relationships that would be intractable or extremely hard to analyze. However, our research philosophy was to describe it as we see it, then worry about how to analyze it.

This insistence on literal description led to the observation of a recurring pattern of functional relationships of the following type:

> If *X* is true, then *Y* is always true; else
>
> If *X* is false, then *Y* may or may not be true, depending on a number of other variables *Z*.

For example,

> If *a shopper does not visit Wal-Mart*, then *the shopper definitely does not buy from Wal-Mart*; else
>
> If *a shopper does visit Wal-Mart*, then *the shopper may or may not buy from Wal-Mart*, depending on a number of other variables such as *Price, Promotion, Service, etc.*

Or,

> If *a shopper does not talk to any salesperson*, then *the shopper definitely was not influenced by the salespeople's knowledge of the products*; else
>
> If *a shopper does talk to a salesperson*, then *the shopper may or may not have been influenced by the salesperson's knowledge*, depending on a number of other variables such as *the salesperson's personality, the customer's own knowledge, etc.*

The defining characteristic of these relationships is that their first half is deterministic: if the `control variable' *X* is set to one value, the value of the `outcome variable' *Y* is instantly determined with complete certainty. The second half of the relationship is stochastic: when the control variable *X* is set to other values, the value of the outcome variable *Y* is unknown, and governed by some probability distribution. In particular, many other variables *Z* causally influence the value that *Y* can take, and one could build a traditional statistical model regressing *Y* on *Z*. Thus the variable *X* acts as a sort of enabling `electrical switch': when flipped off, it cuts off all the other variables *Z* from having any effect on *Y*; when flipped on, it enables the other variables to influence *Z* in their normal probabilistic fashion. Both the deterministic and the stochastic parts are required to fully describe the relationship between *X* and *Y*; hence the name quasi-deterministic function.

These patterns are quite intuitive to understand from a qualitative perspective, as the above examples show. Furthermore, they are surprisingly pervasive, and we found them cropping up all over our model. If you *don't receive an advertising flyer*, its contents can't influence you to *visit a retailer*. If you do receive a flyer it may or may not influence you visit the retailer, depending on what happens next. If you *don't read the flyer*, its contents can't influence you to *visit the retailer*. If you do read the flyer, it may or may not influence you to visit the retailer depending on what's in the flyer. If you hadn't *previously purchased a product from the retailer*, your *feelings about the previous purchase* cannot influence you to *visit the retailer* again (because the feelings are nonexistent). If you did purchase previously, your feelings may influence you to visit or not visit the retailer this time, depending on the nature of that experience. If your *spouse was not present* when you visited a retailer, the *spouse's feelings about the product* could not have affected your purchase decision. If your spouse indeed went shopping with you, it is quite possible that their feelings affected the purchase outcome, depending on how involved they were. If you were not *aware of a promotion* in the store, the *dollar amount of the promotion* will make no difference. If you did notice the promotion, it may or may not influence you depending on your price sensitivity. At a retailer's store, if you *did not seriously consider even one TV* at that retailer, your *will not purchase from that retailer* regardless of all other factors at that retailer; if you did consider at least one TV, all the usual factors such as price, promotions, etc., kick in to influence your final choice of retailer. And so on.

Despite their apparent ubiquity, we have not found these patterns described anywhere in the marketing, choice modeling, or statistical modeling literature. Indeed when we first described these patterns, the statisticians that we consulted advised us to find ways to avoid the patterns and drop them from the model. Their concern was "missing values". If you *did not talk with a salesperson* at a retailer, all the other salespeople-related variables in the model, such as *were the salespeople knowledgeable, were they friendly, did they demonstrate the product, did they help choose a product,* etc., become "inapplicable". When administering the survey to obtain quantitative data, if a respondent indicates that they did not talk with a salesperson, all the subsequent salesperson-questions are skipped. This results in a series of missing data cells (for that respondent) in the quantitative dataset. The missing data cannot be handled with traditional missing-value techniques used in statistics as explained in Section 4.1.4. While the two groups of shoppers (those who spoke to a salesperson and those who didn't) can be analyzed separately from each other, it is hard to combine the results from both groups to draw an inference about the total effect of a variable. When we have numerous such subgroups (and the number of subgroups obviously increases exponentially with the number of such quasi-deterministic relationships) it becomes impossible to do such analyses. As a result, statisticians and survey researchers are loath to allow these `skip patterns' into a model.

The `solution' that was recommended by our statisticians is to change the `troublesome' variables in ways that are meaningful (and enables us to collect data) across both halves of the quasi-deterministic relationship. For example instead of asking "were the salespeople knowledgeable" which is inapplicable when the shopper had not encountered a salesperson, ask "are the salespeople typically knowledgeable" which can be answered if the shopper ever visited the store and spoke to salespeople long ago, even if they didn't talk to a salesperson on the current visit. Indeed an examination of current practice in survey research reveals that most survey questions are phrased using such generalities, and skip patterns are rare.

From a qualitative research point of view, it is obvious that this `solution' is a poor one, because it substitutes precise recall of events that transpired recently by fuzzy general impressions about the past. From a quantitative perspective as well, these generalities have a price. They introduce unwanted noise into the statistical model, because all the things that might have happened during your interactions with retailer's salespeople in the past probably had little effect on your current purchase decision, whereas your current experience probably had a much more direct and greater effect. Thus all the irrelevant previous experiences that are pulled in by the generality of the question are likely to introduce noise in the response data which worsens the fit of the regression model. We hypothesize that one of the reasons for the mediocre explanatory power (R-square) of most market research surveys is due to the large amounts of noise in the questions. Retaining accurate descriptions of natural phenomena such as the examples described above implies that we cannot avoid the use of quasi-deterministic constructs. We believe that doing so will result in eliciting higher quality data and will substantially improve model fit, because the functional form more truly reflects the nature of the underlying causal mechanism that connects the two variables. This is analogous to the manner in which logistic and similar nonlinear regressions improve model fit over linear regressions.

Our focus on capturing variables and their relationships accurately from a qualitative perspective thus created a problem with quantitative modeling. There was no known method for adapting statistical regression techniques to handle quasi-deterministic relationships. Indeed a little reflection reveals the reason for this deficiency. While one half of the quasi-deterministic relationship is a familiar stochastic relationship, the other half is entirely outside the domain of statistics. Indeed most statistical techniques run into convergence issues and other numerical problems,[16] when any cell in the joint probability distribution is strictly zero; indeed many treatises begin with an assumption that such a condition does not exist in the dataset[17]. However, the ubiquity of such patterns, as illustrated in our examples above, repudiates these modeling assumptions.

[16] For example, when we tried a logistic regression on a dataset whose variables were related quasi-deterministically, the coefficient of the control variable progressively increased to very large values and the procedure failed to converge. While this would appear to be a numerical failure of the algorithm, it is actually correct behavior, because the theoretical value of the coefficient should be (plus or minus) infinity. This also induced secondary problems on the estimation of the finite coefficients of the other variables, so a standard logistic regression procedure could not be used.

[17] E.g., see [Pearl 2000, p.15] which requires this assumption to prove the existence of a unique Bayesian network.

We have developed a practical method for quantitative analysis of quasi-deterministic relationships, which we will describe in Section 4.1.4. The availability of this method has considerable significance for researchers, since it enables one to create accurate qualitative descriptions of real world dynamics and obtain a higher quality of quantitative data, without fear of analytical intractability. This is true at least in the domain of human choice behavior where quasi-deterministic patterns appear to be quite common.

4 The Quantitative Modeling Phase

4.1 Methodology for the Quantitative Phase

As introduced in Section 2, one of the primary objectives of the study was to build a robust quantitative cause-and-effect model of the critical factors driving sales. Our research philosophy emphasized a purely exploratory approach, i.e., we wished to start from as clean a slate as possible, minimizing the injection of a-priori assumptions and theories (whether marketing theories of consumer behavior, psychological theories of decision making, or mathematical theories of purchase propensity) into the model. In particular we were mindful of the critiques leveled by cognitive scientists against "folk theories of rational behavior" that assume that shopping processes are goal-directed and therefore presume the common utility-theory based choice models [Lakoff and Johnson 1999 Chapter 23]. Taking such an approach implied that the most commonly used tools of experimental or quasi-experimental research were not of much value to us since we did not have enough prior information about the causal structure of the variables. In particular, confirmatory factor analysis and confirmatory structural equation modeling (SEM), which provide the traditional foundation for such studies in marketing (e.g., as typified by [Novak and Hoffman 2000], were of limited value. Even though we could and did create numerous hypotheses about the relationships between many pairs of variables, there were simply too many variables (over a thousand) and too many possible causal interrelationships between these variables to be able to hypothesize and test a small number of structural equation models. A good quantitative analysis methodology that supports such an exploratory approach has a critical need for mathematical theory and tools that (a) assist with structure discovery, providing automated or semi-automatic discovery of the causal relationships between variables (which can then be tested via traditional structural equation modeling, and (b) assist with disciplined and completely visible introduction of assumptions into the model, whenever trustworthy prior assumptions are available to simplify the model. The latter requirement was considered to be particularly important in our study because we expected that there would inevitably be some causal ambiguities in such a large model (e.g., does variable *A* cause *B*, or is there really an unobserved common cause *C*?). Rather than embed an invisible assumption into the model that resolves the ambiguity but introduces unknown error, we wanted to have the ability to change the assumption and see how much the model changes, i.e., to do sensitivity analyses that indicate the robustness of different parts of the model.

4.1.1 Traditional methods

The two most commonly used quantitative toolkits for such building such exploratory models are the Exploratory Factor Analysis + SEM techniques, and the various data mining techniques such as association rules, Bayesian networks and hidden Markov models. While each of these approaches has some strength, all of them are deficient for building models that respect and reflect the causal structure of the world. It is worth noting that getting causality correct is a very important element of our approach because we had a client who would design business interventions based on our findings, and therefore the need was for *control*, not *prediction.* As Pearl [2000, page 38] explains, for purely predictive models the traditional apparatus of probability theory (regression models, Bayesian nets, etc.) suffices---we fit a model that captures the joint probability distribution of the observed variables, and then for any given value of a variable e.g., *liking-for-brand*, we can predict the resulting *sales*. However, when we make an intervention, such as introducing or discontinuing a particular brand of product, we change the joint probability distribution that governs the modeled variables, and therefore we cannot predict the results in the changed future-world. To make predictions in the changed world, we need to estimate the changed joint probability distributions, which requires knowledge of what aspects of the world are invariant---what hasn't changed between the current and future worlds in the process of making the intervention. In other words, we need knowledge of the causal mechanisms that govern the relevant processes. Any mathematical model aimed at designing interventions and not just predictions---which arguably includes just about every model constructed for business applications, the only exceptions being a handful of purely predictive applications such as stock market prediction[18]---must at least implicitly respect these causal mechanisms, if not explicitly model them. In the following section we briefly review traditional modeling techniques for their consistency with causality, and show that most of them are quite inadequate.[19] We then present the causal modeling methodology that we developed during our study.

[18] Indeed creating an intervention that moves the stock market may result in jail time---this is one of the few situations in which purely predictive models are more desirable than causal models!

[19] If the distinction between predictive modeling and causal modeling were well understood, our critique would be inappropriate because traditional techniques such as Bayesian networks and neural networks [West et al. 1997] were designed for prediction and are well suited for that purpose. However, almost everyone who builds these models recommends some action based on those models, i.e., attempts to derive a causal inference from the technique, which is an inappropriate use. This happens in part because alternative techniques for causal modeling have not been available, and because the difference from predictive modeling is not well known, and therefore predictive techniques are thought to be the only way to draw an inference (outside experimental design). Given the widespread misuse of data mining techniques for that purpose, we feel it is important to review traditional predictive modeling techniques for their (lack of) causal content.

Part of the problem is a difference in philosophy between the causal modeling and data mining approaches. Both attempt to develop pattern-discovery techniques, but most data mining techniques (e.g., multiple regression) are not causally-consistent as we discuss elsewhere in this section. On the other hand, the causal modeling approach seeks to establish causality within the associations found, with the rigor typical of traditional experimental research. This is in contrast with the mining metaphor of hoping to get lucky: if the patterns discovered happen to be causal, the data miner has struck gold; however there is no systematic effort to utilize the semantics of causality in the design of analytical methods. So what if diapers and beer

4.1.1.1 Path analysis and structural equation modeling

Structural equation modeling (SEM) with latent variables is a critical piece of the statistical infrastructure required for building causal models. While SEM has been used in the social sciences and economics, the causal semantics of SEM have been long obscured, with many statisticians refusing to associate causality with SEM models, and practitioners questioning the value of these models [Goertzel 2002]. Pearl [2000, Ch. 5] provides a great introduction and history of the subject, and attributes the problem to an absence within the field of statistics of a vocabulary and calculus for reasoning with causality. In particular, probability theory is symmetric (e.g., consider how Bayes' law relates $P(A|B)$ and $P(B|A)$) and the language of statistics has no means to capture the asymmetry inherent in causality. The unduly complex formulation of SEM concepts such as exogeneity/endogeneity, identifiability, testability, indirect effects, and so on have further served to obfuscate the subject; these concepts are much simpler when formulated using the right causal vocabulary.

In addition to the above theoretical critique, we also take a practitioner's point of view; to us the biggest problem with using traditional SEM is a methodological one: it requires the researcher to guess at an initial structural model, and then guess at modifications that improve the model, without providing much in the way of analytical guidance for model creation and modification. This can substantially destroy the scientific rigor of the model-building process, because two researchers who start with two different theories in mind will create two different initial models and in the end converge to two distinct theories that disagree with each other, but are consistent with the data. The difficulty arises in part because SEM tests of model fit can only reject poor models; however there will always be many models that are not rejected for any set of variables. While it is obvious that external (substantive) knowledge is required to resolve these ambiguities, SEM does not provide a methodological framework for systematically integrating qualitative domain knowledge with quantitative data. It is not an accident that contradictory conclusions are so easily reached as Goertzel [2002] illustrates. The biggest obstacle to spotting these problems when reviewing SEM studies has been the lack of transparency in terms of what modeling assumptions were made and how they were substantiated, which parts of the model are strongly supported, and which parts are ambiguous and uncertain.

While our discussion above refers to the use of SEM in a "confirmatory" mode, in which the research constructs a model *a priori*, SEM has also been used in an "exploratory" mode in conjunction with techniques such as factor analysis. Factor analysis techniques are even more susceptible to the "garbage-in garbage-out" problem described, as we will discuss in Section 4.1.1.2.

Our view is that SEM's current deficiencies can be fixed by (a) switching the vocabulary and computational infrastructure of SEM to directly utilize the language of causal graphs

are commonly bought together? Any useful business application normally requires some causal inference, and the leap from an association to a causal link is left to whims of the data miner. Causal discovery techniques seek to build causal semantics into the algorithms, giving the analyst a basis for making (or not making) causal claims.

and causal calculus as introduced by Pearl, wherever possible, and (b) developing a rigorous methodological framework that allows for reliable introduction of qualitative assumptions, hypotheses, functional forms, and other domain knowledge at the right places in the quantitative modeling process, in a way that greatly increases the transparency of the model's strengths and weaknesses. Given the right conceptual and methodological framework surrounding the mathematics, SEM tools can provide tremendous value not only in traditional tasks such as obtaining statistical parameter estimates and computing effect sizes, but also in supporting semi-automated model discovery. After introducing our causal modeling methodology (Section 4.1.2), in Section 4.1.2.7 we will revisit the topic of how SEM supports and is supported by the methodology.

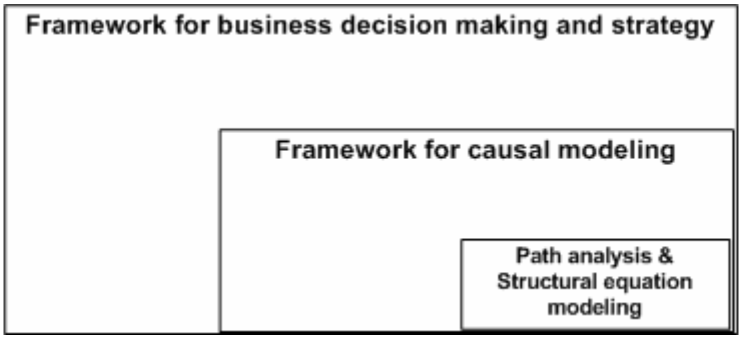

Framework for business decision making and strategy
Framework for causal modeling
Path analysis & Structural equation modeling

Figure 16 Our framework builds upon traditional methods such as SEM

4.1.1.2 Factor analysis, measurement models, and the role of latent variables

One of the modeling approaches that we had originally contemplated utilizing was the traditional exploratory/confirmatory factor analysis (EFA/CFA) methodology [Froman 2001]. In this approach groups of observed variables are treated as `items' which are measures of unobserved `constructs' or `factors'. The relationship between a construct and the items that measure the construct is called the measurement model, and the relationships between factors constitute the structural model [Bollen 1989]. The modeling decision of how to group items together to ensure that the group measures a meaningful construct is either done with the help of discovery algorithms (such as Common Factor Analysis and Principal Components Analysis) in the EFA approach, or guided by prior theory in the CFA approach. In both approaches, the structure of main interest is the relationships between the unobserved factors; the observed items are present in the model merely as a way to measure the factors.

Our focus on establishing causality in models revealed serious defects in these approaches, especially with EFA. The fundamental underlying problem is that the factor analysis algorithms (Common Factor Analysis, Principal Components, or other common dimensionality reduction algorithms used to group items) do not reflect or respect the underlying causal structure between the variables that are being treated as `items.' Since many causal structures can produce the same correlational patterns, it is common for these algorithms to produce factor-item groupings whose implied causality is totally different from the true relationships between the variables. For example variable *A* may causally affect variable *B* which in turn affects variable *C* (Figure 17). As a result all three variables are highly correlated and factor analysis will typically produce a latent variable F which is `measured by' (and is therefore interpreted as the "cause of") all three items A, B and C, and that these three items are independent of each other conditional on F.[20] The resulting erroneous inference would be that that manipulating A will have no effect on B or C, and that one must intervene on F if one wants to induce change in B or C. The erroneous causal conclusions are a serious problem when one is required to use these models to draw business conclusions such as prescribing the best intervention.

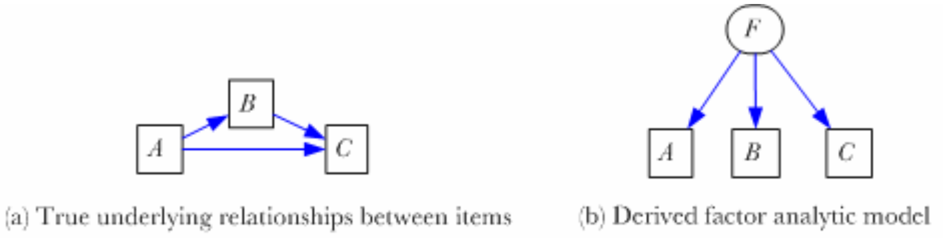

B
A
C
(a) True underlying relationships between items
F
A
B
C
(b) Derived factor analytic model

Figure 17 Causal implications in common factor analysis.
In model (b) variable A does not causally influence B or C, contradicting model (a).

For a real example, consider the construct of "good product selection" which is very important to retail merchandisers because of the common hypothesis that shoppers won't buy from a retailer who doesn't have a good selection of products. Items such as "good selection of brands", "large number of models on display", "wide range of prices", "wide range of TV sizes", etc., are used to measure this construct. During regression analysis it is tempting and indeed common practice to reduce the number of variables by combining all the observed items into a common latent factor, labeled "Selection." The resulting inferences are erroneous in the manner described by Figure 17. "Wide range of TV sizes" is not causally independent of "Wide range of prices" given "Selection"; since increasing TV size often increases price. Similarly, it is perfectly meaningful for a retailer to inquire about the effect of an intervention on "good selection of brands"; if we increase the number of brands that we carry in the store, how much would our sales increase or decrease? A factor analytic model which does not allow a direct intervention on the item "good selection of brands" to have an effect on sales, and only permits an intervention on the latent variable "Selection" to have an effect, is obviously incorrect. Often, reversing the direction of the arrows produces a better match with intuition; the items are the *causes* of "Selection" and sales, and thus the arrows point out of the items towards the factor. The items are then called `effect indicators' [Bollen 1989]. Unfortunately, reversing the arrows violates the assumptions underlying common factor analysis, e.g., that there is a hidden common-cause that *causes* the co-variation of the items, and invalidates the results of the factor analysis algorithms. Confusion about the

[20] In general the items need not be independent of each other conditional on the latent factor; there can be residual correlations between pairs of items. The residual correlations are usually interpreted as being due to other latent common-causes, which implies the same causal conclusion: manipulating item A will not change item C according to the factor analytic model. In general if the causal structure implied by the factor analytic model differs in any way from that of the original items (Figure 17a), the computation of the size of the causal effect of one variable on another becomes biased.

directionality of arrows between factors and items is rampant in the published literature and the cause of a lot of biased studies, as documented by Jarvis et al. [2003].

The problem with causal inference is further aggravated in the EFA approach because the item groupings and factors discovered by these algorithms do not initially have a name or definition, and the researcher is invited to "interpret" these factors. Since the primary structure of interest is the relationships between these latent factors, not the measurement models, there is no way to leave the factors un-interpreted. The ensuing exercise injects a large amount of `creativity' into the model, producing different theories for different researchers depending on their individual prejudices toward their own favored theory, and substantially reducing the scientific rigor of the analysis methodology.

The third problem with traditional factor analysis is the garbage-in garbage-out syndrome, which is unrelated to the causal consistency issues discussed above. This arises from the choice of items that are provided as input to the factor analysis. Invariably, these variables have not been chosen randomly; the researcher has some `inkling' or theory as to what is going on in the domain and has observed data on things of personal interest [Chin et al. 1988, Kirakowski 2001, Igbaria and Parasuraman 1991]. For example, in designing the QUIS instrument [Chin et al. 1988] for measuring the quality of Web sites, the researchers started with a bunch of 90 questions measuring topics of interest such as "learning", "terminology and system information", "screen," and so on. Then principal components factor analysis was performed on the data, and to no surprise, yielded four latent factors called "Learning", "Terminology and information flow", "System characteristics" and "System output" as the major factors driving website usability. Since the process of choosing which items will be measured by a questionnaire is implicitly guided by the concepts and theories already in the researcher's mind, it is inevitable that during analysis these measures simply get regrouped into factors that correspond precisely to those preconceived concepts. Thus factor analysis is extremely weak as a discovery tool---it mainly discovers what was in the researcher's head, not the true structure of the phenomena being observed.

The last criticism can be significantly mitigated by using a better item-selection process, e.g., by using robust theory-less qualitative research methods. However the earlier two criticisms basic on causal grounds imply that current factor analytic techniques are quite inadequate for introducing latent variables into structural equation models in the hope of gaining model parsimony. They need to be replaced better techniques for detecting the existence of latent variables that that could simplify the relationships between large numbers of observed variables, and these techniques must work with many different underlying structures (e.g., causal arrows going from the items to the factors), not just the "common-factor" structure.

One such causally-consistent technique is the method of Vanishing Tetrads [Shipley 2000, Spirtes et al. 2000]. This technique searches for groups of four variables whose pairwise correlations cancel out in a tetrad test, e.g., $\rho_{AB} \cdot \rho_{CD} - \rho_{AC} \cdot \rho_{BD} = 0$. The greater the number of vanishing tetrads associated with a group of variables, the greater the indication that there must be a latent common-cause causing the co-variation within that group. When such latent variables are identified, we generally have no basis for naming and interpreting them, because the source of co-variation may be quite remote from the actual observed variables, and there may be more than one reason why the observed variables co-vary (i.e., the latent variable is actually standing in for a group of unobserved mechanisms). However, discovering the existence of these variables can yield substantial simplification of the causal links between the observed variables, e.g., replacing a complete graph by a star-shaped graph. Speculating on the interpretation of the latent variable may not help the current study, but provides insight for guiding a future study in search of the common causes.

In conclusion, common factor analysis and similar dimensionality reduction techniques appear to be useful only for building purely predictive models, from which no causal conclusions or business interventions will be drawn. When business actions have to be drawn from these models, it is extremely important to ensure that the modeling techniques used are at least causally consistent, i.e., do not permit the drawing of inferences that violate underlying causal structure, even if the techniques themselves do not explicitly incorporate causal modeling concepts. Tetrad-like analyses hold some promise in this direction [Shipley 2000].

As a result of our experiments with these methods, we dropped factor analysis from our research methodology. We experimented to a limited extent with tetrad algorithms to obtain guidance on where to insert latent variables, but found few clear markers. In a few instances, we relied on theory. In the end however, our strong reluctance to introduce fictitious variables with little theoretical support resulted in very few latent variables being used in the model. As far as we can tell the results are excellent: the model structure discovered by our methodology had good fit, and the interventions and causal effects we computed were completely meaningful and appropriate. We hypothesize that model fit would have been improved even further if we had introduced more latent variables because of the reduction in measurement error, but until we develop better causally-consistent algorithms that provide guidance on where to insert such variables and how to interpret them in a scientifically rigorous manner, we recommend the more conservative but reliable approach described above.

4.1.1.3 Bayesian networks and hidden Markov models

Bayesian networks are similar to causal networks; indeed if one ensures that the arrows in a Bayesian network are consistent with the direction of cause-and-effect between the corresponding variables, and if one superimposes the calculus of causality, a Bayesian network becomes a causal graph. The catch of course is that Bayesian network algorithms have no way to ensure that the directionality of links is correct. If the modeler happens to set the arrows in the correct direction, and if inferences are drawn utilizing the causal calculus (e.g., computing $P(Y|\text{do}(X))$ instead of $P(Y|X)$) the results will be correct, but if the modeler's structural knowledge is not reliable the conclusions will be erroneous. Unfortunately, it is all too easy to permit the arrows to point the wrong way, given the intrinsic symmetry of probability theory, unless causality detection methods are introduced. Therefore interventions derived from traditional Bayesian networks will always be suspect.

Since hidden Markov models are a special form of dynamic Bayesian network, the above comments apply to those models as well. If the structure of the HMM has been correctly predetermined, say with reliable expert knowledge, the rest of the estimation will produce the correct results. On the other hand if the structure is unknown a priori and must be inferred from the data, causal modeling techniques have to be introduced.

4.1.1.4 Regression trees

The problem with simple multivariate regression is relatively obvious: one cannot estimate the (total) effect of an explanatory variable on the dependent variable unless one knows the interrelationships (full causal structure) between the explanatory variables. Assuming that all the explanatory variables are a priori known to be causes of the dependent variable, and assuming that all confounding common-cause variables are present in the regression---the second assumption being especially difficult to justify---the best that can be said is that the direct effect of each explanatory variable is obtained via the regression. In such a case one cannot compute the effect of making an intervention on any one of the explanatory variables, because the interpretation of a regression coefficient as an effect size requires that all other explanatory variables be held constant, which in practice is an extraordinary requirement; see our discussion of this topic in Section 4.1.3. This is not a useful intervention, and simple regressions do not allow us to compute the results of practical interventions; a full structural equation model is needed for that.

Regression tree methods such as CART and CHAID try to go further by uncovering additional structure between the explanatory variables. The problem is that uncovering such correlational structure still does not solve the problem of resolving causal direction. If the variables are organized into a tree, what gives us the certainty that cause and effect go up from the bottom of the tree to the top? Even if the tree arrangement happened to coincide perfectly with causal relationships between the variables, are the relations between the variables truly tree-like and not a graph? I.e., are the implied conditional independence relationships truly justified? As with any structure deduced via purely correlational methods devoid of causal checks, these methods produce models whose (causal) correctness is unknown, and therefore we cannot rely on them to derive business interventions.

4.1.1.5 Value decomposition trees and value networks

Keeney [1999] and others have utilized various tree or network based decompositions to model the relationship between decision makers' "goals" or "values". For example a means-ends decomposition can be constructed via laddering techniques [Reynolds and Whitlark 1995] to identify the variables that ultimately achieve the shopper's goal of making a purchase, or to describe the utility of the purchase. These models generally have the problems identified in the previous section on regression trees: if we ask what is the effect of an intervention on a leaf node of the tree or at any point in the network, it is hard to derive a defensible estimate of the resulting improvement in goal achievement or the increase in value. The key point that modelers need to keep in mind is that decomposition of something into its constituent components does not necessarily correspond exactly to the flow of cause and effect.

4.1.2 Our Causal Modeling Methodology

We have earlier identified several weaknesses with traditional SEM and other quantitative methods, including the lack of a good conceptual vocabulary and calculus for reasoning about causality and drawing causal inferences; the lack of automated support for discovering the correct model structure from quantitative data without relying on *a priori* theories; the inability to scientifically integrate domain knowledge obtained from ethnographic and other qualitative research; the lack of transparency into a model's regions of strength and weaknesses; and the lack of guidance for model improvement.

We have also indicated that applying the theoretical concepts and algorithms developed in Pearl [2000] and Shipley [2000] is an important first step towards solving these problems and developing a better quantitative methodology.

However there are many practical issues that have not been solved, since causal modeling is a very young field of research. Experimental tests of the algorithms on real data have usually shown poor or mixed results, and many experienced statisticians are skeptical of the ability of causality-detection algorithms to extract useful structure from data [Freedman 1993, 1998]. The problems that have been solved on real data have usually not been much larger than toy problems. As model size increases, the likelihood of errors in the discovered model structure goes up. Only one software tool has received (relatively) wide attention and testing, namely Tetrad [Spirtes et al. 2004]. We obtained a copy of Tetrad IV and tested its performance on several groups of variables from our quantitative dataset, and found that the models produced were incorrect, i.e., they inferred causal directionality that we knew to be impossible from our substantive knowledge of the variables. The Tetrad software is known to require tweaking of several parameters by the user to avoid producing nonsense models, and it is possible that we did not tweak the parameters adequately. However, the tweaking itself is a methodological problem since there is no way for the user to know what parameter settings produce a correct model, and which don't. Also, if the user has enough substantive knowledge to detect errors in the models produced, tweaking of parameters whose precise influence on the various parts of the model is unclear, is not the best way to take advantage of that domain knowledge. The modeler should be more intuitive controls over the performance of the discovery algorithms.

Our examination of the theory behind the algorithms revealed theoretical problems as well. In particular, we discovered that the statistical tests used to extract independence-structure appear to be incorrectly designed; also, the algorithms do not handle ambiguous situations very well and are too aggressive in making causal inferences under some circumstances. Because of the iterative nature of the algorithms, one incorrect causal inference (e.g., wrongly setting the direction of one link in the model) can get seriously compounded and subsequently produce a whole bunch of incorrect inferences. Since the algorithms involve a lot of computation, software such as Tetrad does not allow the user to catch the first error, manually prevent it, and let the rest of the inferences go on correctly; instead the user is faced with a model with a large number of errors in the structure and no idea how to correct it. Thus designing traceability of every step into the software design, maintaining a comprehensive history, allowing the user to change any inference in the sequence and instantly recomputing subsequent inferences is crucial for handling the inevitable errors that occur when creating large models in the presence of statistical noise.

Perhaps our most important empirical discovery with theoretical significance is that the vast majority of the errors produced by the causality detection algorithms turned out to be due to violations of the "stability assumption" (aka the faithfulness condition or Causal Markov condition) required by these algorithms.[21] The algorithms provided by Pearl, Spirtes, et al., and the Tetrad software, provide no recourse under violation of this assumption. On the contrary they believe violations of this assumption are rare, and provide various theoretical defenses of its improbability [Spirtes et al. 2000].

However, we found that the algorithms often produced spurious cancellations, which are violations of this assumption. The prototypical examples of these violations is shown in Figure 18, in which *B* should never be independent of *C* given *D*, since conditioning on the collider *D* activates a correlation between its parents *B* and *C*, and the common-cause *A* (or a direction cause from *B* to C) induces another correlation between *B* and *C*. Unfortunately, we often found $B \perp C \mid D$. The explanation turns out to be quite simple: most of the links in our models are usually positive in nature (e.g., when *B* increases, *D* increases), so the correlations induced by the upper paths are positive. Conditioning on a common-effect *D* induces a *negative* correlation between *B* and *C*. This results in at least a partial cancellation of correlations, and increases the likelihood of finding $B \perp C \mid D$. Testing for another independence when conditioned on two variables like *D* that are common-effects of *B* and *C* further increases the likelihood of cancellation. If we do enough statistical tests for zero correlations, sooner or later one of the tests will produce an error. For example, at 95% confidence levels, roughly 1 in 20 tests will produce an incorrect inference purely due to statistical error. Since a model with 10 variables will have at least 360 such tests of independence conditioning on one variable, it is inevitable that even tiny models are susceptible to purely statistical error and incorrectly conclude that the link between *B* and *C* should be dropped (during the first stage of these algorithms, when independence-structure is computed). Since dropping a link when a correlation actually exists is a far more serious error that retaining a link where there is no correlation, and because an erroneous inference is propagated to the rest of the model during subsequent stages of the algorithms, the problem quickly gets compounded to a serious degree. When an error is spotted by the modeler after much computation downstream, the modeler is often quite confounded about the source of the problem, since the error may have been noticed in some other distant part of the model.

[21] Since we discovered this on our consumer behavior dataset but have not examined errors in analyzing datasets from other areas, e.g., from biology, we do not know if this phenomenon is peculiar to the domain, i.e., it may be a characteristic of human behavior but not true of other kinds of systems. We hypothesize that it is more general, i.e., that examination of errors in causal models derived from datasets in other domains will also turn out to be usually violations of the Causal Markov condition, and the same fixes that we describe will resolve the problems.

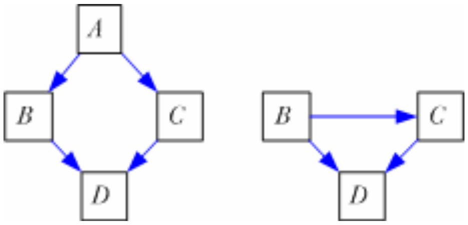


Figure 18 Spurious cancellations: common patterns that produce violations of the stability assumption

The realization that the primary causes of error in existing causality-detection algorithms is due to the spurious cancellation patterns of Figure 18 is quite encouraging because knowing the nature of the error permits designing better algorithms that correct for the errors. The really good news is that whenever these errors happened to us, they were easily detected downstream during the model estimation step by SEM tools, and the modeler could in the span of a few minutes isolate the root cause by visual inspection, despite the number of subsequent inferences performed by the model. Because of the traceability features of our Causal Modeling Workbench, a modeler could immediately detect that the link between *B* and *C* was incorrectly dropped due to "conditioning on an effect". This judgment was made using domain knowledge where the modeler knew that *D* could not be a cause of *B* or *C*---which on the surface appears to require substantial knowledge and assumptions on the part of the modeler, but in our experience turned out to be surprisingly easy to recognize and defend, without feeling that some unjustified assumptions were being introduced into the model. Note in particular, that the solution was to reintroduce a link between *B* and *C*; we did not actually introduce any substantive assumptions into model, but merely removed an assertion of independence inferred by the algorithms. This weakened model was then reprocessed by the algorithms as before, and ended up producing a completed model that was judged to be correct at least in the sense that they did not violate any domain knowledge.

Thus, modification of the algorithms of Pearl et al., to detect and correct spurious cancellations, handle ambiguous conditions, use better statistical tests, and introduce qualitative assertions systematically produces a much-improved methodology and analytic toolkit from a practical implementation point of view. Our belief is that these improvements will substantially increase the real-world effectiveness of causal modeling methods, and go a long way towards overcoming Freedman's critiques and the general skepticism of many statisticians about the practical usefulness of these methods.

Since causal modeling algorithms cannot be used in a fully-automated manner because of need to insert substantive assertions at the right places, they need to be conceptualized in terms of a larger framework and methodology for interactive modeling with software tools. We developed our own toolkit (aka the Causal Modeling Workbench), in the form of a collection of algorithms and utilities programmed on top the statistical analysis platform R, and interfaced to SEM tools such as Mplus. The workbench can take a quantitative dataset as input, detect as much causal structure as possible from the data, accept qualitative assertions from the modeler (starting with no assertions, then slowly increasing the number of assertions used, introducing them in order of robustness) till just enough assumptions have been made to complete the model structure, and generate an Mplus specification of the resulting SEM for estimation. Model The workbench provides support for tracing inference history, as well as controls for trading off the risk of making statistical errors against model complexity.

This larger framework for quantitative model development is important because it helps address the problems identified at the beginning of this section. Figure 19 illustrates the main elements of the framework. The guiding principles that led us to order the steps in the sequence depicted were (a) automate as much of the modeling process as can be reliably automated given the best algorithms that we have, and (b) defer the remaining manual manipulation of the model as much as possible in order to minimize the impact of and biases from the modeler's assumptions. Phase 1, automated causal discovery, relies on our variant of Pearl's IC* algorithm, which we dubbed the IC+ algorithm, to extract as much structure as can be reliably extracted before the modeler is asked for substantive assumptions in Phase 2. The assumptions obtained in Phase 2 are use in a minimalist manner: since each assumption has the possibility of enabling automated discovery of additional causal structure, we first apply an assumption, then reiterate with the algorithms of Phase 1 to extract as much structure as possible, then apply another assumption from Phase 2, and so on. After the entire causal graph has been constructed, Phase 3 generates the induced structural equation model and estimates the model parameters that provide the best fit with the data. Correcting problems with model fit can sometimes trigger backtracking to change modeling decisions made in Phases 1 and 2, but this is surprisingly minimal as explained in Section 4.1.2.5. The finished quantitative model is then used for scenario development and impact analysis as described in our overall framework (Figure 5).

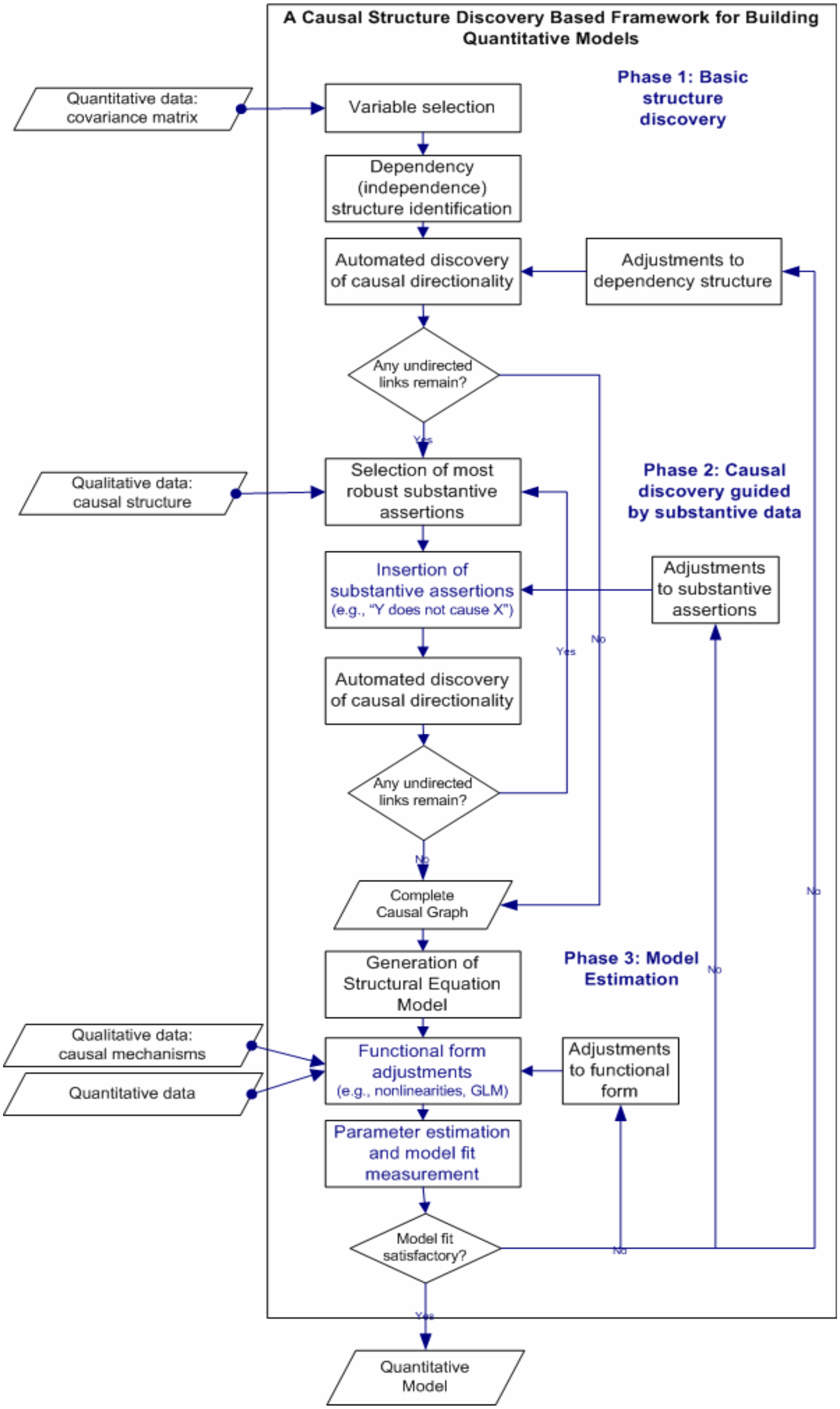

A Causal Structure Discovery Based Framework for Building Quantitative Models
Phase 1: Basic structure discovery
Quantitative data: covariance matrix
Variable selection
Dependency (independence) structure identification
Automated discovery of causal directionality
Adjustments to dependency structure
Any undirected links remain?
Yes
Phase 2: Causal discovery guided by substantive data
Qualitative data: causal structure
Selection of most robust substantive assertions
Insertion of substantive assertions (e.g., "Y does not cause X")
Adjustments to substantive assertions
Yes
No
Automated discovery of causal directionality
Any undirected links remain?
No
Complete Causal Graph
No
Phase 3: Model Estimation
No
Generation of Structural Equation Model
Qualitative data: causal mechanisms
Quantitative data
Functional form adjustments (e.g., nonlinearities, GLM)
Adjustments to functional form
Parameter estimation and model fit measurement
Model fit satisfactory?
No
Yes
Quantitative Model

Figure 19 The Quantitative Phase of our decision making framework

4.1.2.1 Discovering independence-structure

While the IC* algorithm [Pearl 2000] provides the theoretical basis for the discovery of causal structure, its empirical performance on real data has been problematic. We found that our direct implementation of the IC* algorithm generated models that were inconsistent with our causal understanding of the variables. The same problem occurred when we ran Tetrad on our dataset. Tracing the errors back through the algorithms to their root causes showed that these algorithms were being too aggressive in drawing causal inferences, e.g., asserting that two variables were (conditionally) independent of each other in a borderline situation where it would have been wiser to refrain from drawing this inference. Further examination revealed that there were two fundamental underlying problems with these algorithms: (a) the binary nature of the inferences: when testing the relationship between two variables, the algorithms either concluded that the variables were independent or dependent; they couldn't conclude that the variables "may be independent," and handle the resulting ambiguity properly; and (b) misuses of hypothesis testing procedures: the statistical test that was used to infer independence, viz., the Fisher Z test for zero correlations, is actually designed to infer dependence. In other words, if the correlation is so large that Fisher test rejects the null hypothesis (that the two variables are independent), we can correctly infer that the two variables must be dependent; however if the correlation is somewhat small and the Fisher test does not reject the null hypothesis, we should not infer either independence or dependence. Unfortunately, the standard implementations of the causality algorithms use a negative result from the Fisher test to infer independence. This leads to incorrectly dropped links in the causal graph, which sets off a chain of incorrect inferences. Indeed the misuse of this test has led to pathological recommendations in the literature on how to set confidence levels appropriately: for small sample sizes, we are advised to use a more relaxed (lower) confidence level to infer independence [Shipley 2000, p.284]---in other words, given noisier data we are asked to be more accepting instead of more suspicious about drawing inferences correctly![22]

[22] Since the Fisher Z test checks whether the correlation is significantly different from zero, and since the becomes more sensitive as sample size increases, a correlation of 0.01 may be end up as insignificant when *N=100*, and the same value significant at *N=10000*. In fact, any miniscule correlation can be treated as being `significant' given a sufficiently large sample. The result is that as *N* increases, the causality detection algorithms are unable to find any independence relationship, and therefore unable to drop links in the causal graph or detect any causal structure. This is clear evidence of incorrect algorithm design: as the sample size increases, we are getting more and more information about the underlying phenomena, which should put us in a better position to draw correct inferences about the relationships between variables. It is pathological to argue the reverse: that reduced sample size is `better' because it enables the algorithms to detect more structure, which either means we should throw out good data or we should not accept a dependency despite detecting it at a 95% confidence level, and we should artificially keep raising the bar until we get "enough" structure. These methodological peculiarities are symptoms of an underlying weakness in the statistical approach: the knobs being controlled (cutoff levels) do not correctly map to overall statistical error. If two variables are indeed independent in the real world, the greater the sample size of our observations, the more surely we should infer that they are independent---a property that does not hold true if we use the Fisher test for zero correlations as in Tetrad [Spirtes et al. 2004] and Shipley [2000].

4.1.2.1.1 Testing for independence

To design a statistical test that properly infers independence, we need to set up an inverted null hypothesis ("the two variables are dependent") and then reject that hypothesis at the desired confidence level. In other words, instead of testing for a zero correlation, we have to test for a nonzero correlation, and show that the probability of the correlation being nonzero is small enough to justify inferring independence. Unfortunately this makes the statistical test dependent on the particular size of the correlation: how large is "nonzero"? The situation is analogous to the calculation of power in hypothesis testing; the power of an experiment depends on the size of the effect that must be detected. In other words, we are forced to specify the equivalent of an "effect size", in this case a correlation size, that allows us to state that the relationship between the two variables is "negligible" enough to be classified as independent when a suitably designed statistical test reject the null hypothesis. Thus the restated null hypothesis is of the form "the correlation is at least 0.1 in magnitude"; rejecting this with 95% confidence gives us the license to infer that the correlation is small enough to justify declaring the two variables independent.

This approach immediately raises two questions: how should we pick the cutoff value of the correlation that constitutes a non-negligible dependency, and how do we test for it? The answer to the first question is an empirical one (also see below for a theoretical view): research studying real-world applications of hypothesis testing has shown that the best practical definition of a "trivial" correlation is anything below 0.1 in magnitude; correlations between 0.1 and 0.3 are considered "small", and so on [Hopkins 2000, Cohen 1988]. Thus we use 0.1 as the default value, and software implementations allow this to be adjusted in a more or less conservative direction as desired, trading off model simplicity against accuracy.

The answer to the second question is a variant of Fisher's Z test: the test can be modified to examine whether a correlation is within a particular range of values, instead of testing for a zero value.

Given any normally distributed population *N(m,s),* where *m* is the mean and *s* is the standard deviation, and given that our decision rule is to infer that the mean is 0 (negligible) if the observed mean value is within the interval [-x_c, x_c] where the cutoff x_c is a positive quantity, probability of misclassification using this decision rule (i.e., the probability of concluding that the true mean μ is inside [-x_c, x_c] when μ is actually outside that range) is given by $p_e(m,s,x_c) = 0.5\left(\operatorname{erf}\left(\frac{|m|+x_c}{s\sqrt{2}}\right) - \operatorname{erf}\left(\frac{|m|-x_c}{s\sqrt{2}}\right)\right)$, where erf is the error function. Since applying the Fisher transformation $F(r) = 0.5\log\left(\frac{1+r}{1-r}\right)$ to the Pearson correlation r results in a Z distribution whose standard deviation is $S(N,V) = 1/\sqrt{N-V-3}$ where N is the sample size, and V is the number of conditioning variables (needed when testing a partial correlation; *V=0* for a bivariate correlation), we can therefore compute the error probability as $p_e(F(r), S(N,V), F(r_c))$ where r_c is the desired correlation cutoff (with default value 0.1), and *F(r_c)* is the Fisher-transformed value of the correlation cutoff. To standardize this into the usual *p*-value format, we need to use 1 – p_e instead of p_e. Therefore if $1 - p_e \geq 0.95$ we can safely conclude that given the observed correlation *r*, the true correlation must be within the range [-r_c, r_c], and is therefore small enough to concluding that the variables involved must be independent.

We called our above variant of the Fisher Z test, the "Fisher Z with finite ρ" since we are testing for nonzero correlation values. Graphical exploration of our error probability as a function of the observed correlation *r* and sample size *N* is insightful. It reveals that our test is more conservative when the correlation value is around 0.1, but performs identically to the basic Fisher Z test used by Tetrad whenever the observed correlation is large (>> 0.1) or very small (<< 0.1), since a very large correlation is strong indication of dependence, and a correlation extremely close to zero is strong indication of independence, regardless of the sample size (so long as the sample size is not trivially small). In particular, note that for very small correlations, increasing sample size makes us increasingly certain that we can infer independence, which is in distinct contrast to the behavior in Tetrad [Spirtes et al. 2004], Shipley [2000], etc. Around an observed correlation around 0.1, our test becomes sensitive to sample size. With a large sample the standard error of the correlation is small, and therefore if the correlation is slightly smaller than 0.1, our test infers independence; at a slightly higher value, the test infers dependence. This is a natural result of our decision to use 0.1 as the `clinical threshold' [Hopkins 2000] between relationships that we drop and relationships that we keep in the model. Also note the discussion in the next section about the intermediate "ambiguous zone".

We feel there is an additional theoretical justification to this approach. In complex real world models, it is unlikely that we would find two variables that are perfectly independent of each other (conditioned on some other variables). While it is reasonable to imagine that the price of tea in China should be independent of the amount of traffic in Los Angeles, and while this would indeed be well-justified as a substantive assertion according to the principle of "no action-at-a-distance", it is hard to justify taking an extreme absolutist position that the correlation must be strictly zero. The more realistic statement is that any correlation would probably be so small that it can be treated as zero for practical purposes. In other words, most of the independence relationships that we assert in order to keep the model simple are a result of `rounding off' or declaring the relationship `sufficiently negligible' that the resulting model would still be correct for practical purposes. Precisely what size of relationship constitutes "negligible" is an important modeling decision that ought to be explicitly addressed by the modeler, since it affects the accuracy of the resulting model estimates. There cannot be an absolute basis for picking one cutoff size or another; instead, giving the modeler control over this decision based on their desire for model simplicity and tolerance for inaccuracy appears to be philosophically the right approach.

4.1.2.1.2 Managing ambiguity

While our "Fisher Z with finite ρ" test improves upon the basic Fisher test in terms of correctly formulating the hypotheses being tested, there is still the issue of ambiguous inferences. Specifically, if our test shows that $r \in [-0.1, 0.1]$, $1 - p_e \geq 0.95$, we can properly infer independence, but what if $1 - p_e = 0.7$? Should we immediately conclude that the underlying variables must be dependent, since dependence is the conceptual opposite of independence? If $1 - p_e < 0.05$ we can reasonably conclude that $r \notin [-0.1, 0.1]$ and therefore we can infer dependence. However the intermediate *p* value of 0.7 is problematic since it does not fully support either an inference of dependence or of independence. This is an ambiguous situation, and in our view, the best way to deal with such ambiguities is to make the most conservative inferences possible in the circumstance. In the first step of IC* (and our variant IC+), the conservative action is to preserve links between variables (because a link can eventually be estimated with a zero or small coefficient and then dropped), whereas dropping a link is dangerous (because the absence of a link triggers additional inferences in subsequent steps of the algorithm, not to mention forces the structural parameters to zero). Therefore the correct action is to retain the link and proceed with the algorithm. Note that retaining a link in the causal graph is *not* interpreted as *assuming* dependence between the variables, but merely *permitting* dependence.

Managing this zone of ambiguity---in which we can neither infer dependence or independence from the results of the statistical test---takes on much greater significance during the subsequent steps of the IC algorithms, in which we detect colliders and causal chains. The reason is that these more complicated patterns require tests of independence over multiple links at the same time, and require one test to indicate dependence while another test simultaneously indicates independence on the respective links. A conservative approach therefore requires $1 – p_e$ to be below 0.05 or above 0.95 on the respective tests; if either test falls in the range [0.05, 0.95] we have an ambiguous situation where we do not have enough evidence to prove that we have a collider or chain. Our IC+ algorithm modifies the steps of IC* to handle ambiguous situations correctly.

4.1.2.1.3 Testing nonlinear relationships

The use of Pearson (linear) correlations as a measure of independence can be risky when relationships between variables are highly nonlinear. We experimented with nonparametric (Spearman) correlations as well as polyserial and polychoric correlations (for ordinal variables), and found the results to be substantially similar to using Pearson, at least for our application domain. Our view is that even when the relationships between variables are nonlinear, if the relationships are monotonic, the Pearson correlation is in practice likely to provide a good basis for testing independence structure. However, when severely nonmonotonic relationships are suspected it may be advisable to transform the variables suitably before applying causality detection methods such as IC+. Using generalized correlations [Idé 2005] is also another possibility, which we did not explore.

Since the statistical tests for independence depend on sample size, quasi-deterministic data (Section 4.1.4), which reflect a special type of nonlinear relationship, poses an additional difficulty: when data on some variables is systematically missing, the effective sample size is reduced. Of course, if we are testing the correlation between two variables that have been quasi-deterministically switched by the same control variables, then we would have identical amounts of data on both variables (apart from the small amounts of randomly-missing data), and the effective sample size is easily identified. However, if one of the variables is unswitched, and the other is quasi-deterministically missing, the effective sample size can be treated to be the size of the latter variable (i.e., the smaller of the two) in a conservative approach. However, this underestimates the true effective sample size, since the missingness is systematic---an issue that becomes more significant when we replace the latter variable by its reconstituted or switched-product versions (Section 4.1.4.3) as required for causal-modeling. In the case of a reconstituted variable, the effective sample size is again straightforward: it is the full sample size (after the missing values have been restored), since the variable has independent meaning on its own. However, the switched-product version of the variable does not have independent meaning; it is a proxy variable that can be correctly used in place of the original variable regressions and covariance matrices (Sections 4.1.4.3 and 4.1.4.6), but it really does contain less true information despite possessing data in all cells. So the tests for independence must be adjusted to reflect the lower sample size.[23]

4.1.2.2 Discovering directional structure: the IC+ algorithm

While the previous step started with an undirected complete graph and ended with a graph in which as many links as possible were dropped, the next step produces a partially oriented graph, where arrowheads have been added to any link where causality can be automatically established. Our adaptation of Pearl's [2000] IC* algorithm is presented in Figure 20. One visible enhancement is breaking up Step 2 into two explicit statistical tests, 2a and 2b; the original IC* algorithm had only one test, viz., 2a. The rationale behind this breakup is the following. To prove that three variables *a*, *b* and *c* are in a collider relationship with each other (*a* -> *c* <- *b*), we need to establish two conditions: (1) $a \perp b \mid S_{ab}$ and (2) $a \perp \; b \mid c, S_{ab}$ where S_{ab} is some set of conditioning variables that does not include *c*. The first condition requires a test for independence; the second a test for dependence. In the original IC*, since dependence is automatically assumed if independence is not proved, the second test was redundant as it could be inferred from the previous tests (in Step 1). However, since we use a 3-valued characterization of each link, independent, dependent, or ambiguous (the last situation arising when the statistical test for independence cannot conclude with high confidence (p > 0.95) that the correlation is either small or large, as described earlier), we cannot treat a failure to prove independence as proof of dependence. Thus we need to introduce Step 2b into the algorithm, where we explicitly check for the ambiguous case. If there no ambiguity, we

[23] Our software code was not programmed to make this correction, since it lacked full support for representing quasi-deterministic data. This may have led to more aggressive inferences of independence in some of the switched subsections, inducing greater structure (link deletions) than would be warranted in those sub-areas. However, there were no adverse effects apparent to us when we visually examined the models; further, goodness-of-fit tests of the induced SEMs were excellent. Nonetheless, we recommend making the corrections described for QD data; permitting overly aggressive inferences from an inflated sample size has the risk of creating model errors that cannot be easily corrected by tweaking an overall control parameter (such changing the definition of a `small correlation' to some value greater than 0.1, or tightening the p-value acceptance level to something greater than 0.95) since the errors affect only the switched variables, not the entire dataset.

conclude that the three variables are in a collider relationship, as in IC*; however if there is ambiguity we take a more conservative approach and draw no inference.

Figure 20 The IC+ algorithm for causal discovery

1. For each pair of variables a and b, search for a set S_{ab} such that $a \perp b \mid S_{ab}$. If there is no such set, draw an undirected link between a and b.
2. For each pair of nonadjacent variables a and b with a common neighbor c,
   For each S_{ab}, check whether $c \in S_{ab}$.
   a. If $c \in S_{ab}$ continue to the next S_{ab}. (This S_{ab} cannot be the basis for establishing a collider pattern. Record "noncollider".)
   b. If $c \notin S_{ab}$, check whether $a \perp b \mid c, S_{ab}$
      i. If so, continue. (Ambiguous case: may or may not be a collider. Either or both of the a-c and c-b relationships, after partialing out S_{ab}, is likely to be weak, and when combined into the a-c-b chain, the resulting partial correlation $a \perp b \mid c, S_{ab}$ therefore stays small.)
      ii. Else, we have a dep $b \mid c, S_{ab}$ Add arrowheads pointing at c, i.e., a -> c <- b. (Record Collider at c.)
3. In the partially directed graph that results, recursively add as many arrowheads as possible according to the following rules:
   a. For each pair of nonadjacent variables a and b with a common neighbor c, if the link between a and c has an arrowhead into c, and if the link between b and c has no arrowhead into c:
      i. Check whether $c \in S_{ab}$ for any of the S_{ab}.
         If not, (i.e., if none of the S_{ab} contain c), continue. (Ambiguous case. *Proof*: In the previous step, $c \notin S_{ab}$ implies either Ambiguous or Collider. Collider is now ruled out by the absence of an arrowhead from b to c.)
      ii. For each of the S_{ab} that contains c,
         - Check whether $a \perp b \mid S_{ab} - \{c\}$. If so, continue. (Ambiguous case; results from S_{ab} being larger than needed to block a-b, i.e., c could have been dropped from S_{ab}. This case won't be encountered if S_{ab} was the smallest blocking set.)
         - Else, we have a dep $b \mid S_{ab} - \{c\}$ for some S_{ab}. that contains c. So add an arrowhead on the link between c and b and mark that link to obtain c -*-> b (Definite cause).
           If such an arrowhead is found, terminate this substep, skipping any remaining unchecked S_{ab} that contain c.
   b. If a and b are adjacent, and there is a directed path composed of strictly marked arrows from a to b, then add an arrow into b on the link from a.

A similar change is made in Step 3a, where the search for definite-cause relationships is handled conservatively by adding an extra test to identify the ambiguous case.

There is another, more subtle, implementation change that is not shown in Figure 20, but which has a big impact on the accuracy of the algorithm. As mentioned earlier, it is quite common to make inference errors (e.g., to conclude $a \perp b \mid S_{ab}$ when really $a \perp b \mid S_{ab}$) because of the sheer number of statistical tests executed on even a moderately large set of variables (a few dozen). This may be the cause for another empirical problem that we observed: at the end of step 2 of the algorithm, which searches for colliders, a triplet *a—b—c* may have been labeled both a "noncollider" and a "collider". The two labels correspond to two different sets of conditioning variables; one set S_{ab} may satisfy the conditions of Step 2a resulting in the triplet being labeled a noncollider, and another set of variables S_{ab}*'* may satisfy the conditions of Step 2b(ii), resulting in the triplet being labeled a collider. Since the algorithm exhaustively searches through all combinations of conditioning variables S_{ab} taken one at a time or two at a time[24], a 50-variable model requires over 1200 tests for each triplet, and the chance of statistical error is nontrivial. Since a triplet cannot logically be a noncollider and a collider at the same time, and unfortunately we do not know how to tell which of the two labels is incorrect[25], our response to such contradictory inferences is again to be conservative and refuse to conclude that the triplet is a collider, unlike the less conservative IC* algorithm. This approach requires us to implement second step of the IC+ algorithm in two passes (Figure 21).

Figure 21 <u>A two-pass version of step 2 of the IC+ algorithm</u>

Pass 1 (Detection): For each pair of nonadjacent variables *a* and *b* along with a common neighbor *c*, apply the logic described in Figure 20, Step 2, to label the triplet a noncollider or collider. Do not actually modify the underlying causal graph to add arrowheads; just produce labels.

Pass 2 (Acceptance): For each triplet, examine its set of labels. If one of the labels indicates that the triplet is a collider, check whether any of the other labels says "noncollider". If no such inconsistency is found, go ahead and treat the triplet as a collider, modifying the causal graph with the appropriate arrowheads. Otherwise, continue without any action.

[24] We intentionally restricted the number of conditioning variables to not exceed two, because the risk of erroneous inference keeps rising as the number of such tests increase. We found that a tighter restriction (one variable) significantly reduces the number of independences found, and a looser cutoff (3 or higher) significantly increases errors; 2 appears to be an empirically good compromise.

[25] We can consider using a voting scheme: after all the tests are done, count the number of times each label appears. However, we did not use this heuristic as we do not know its effectiveness; simulation experiments are required.

4.1.2.3 Injecting theory into the model

When given just a quantitative dataset and no substantive information, a fully automated run of the IC+ algorithm typically discovers a limited amount of causal structure. It is excellent at discovering dependency structure, typically eliminating a third to two-thirds of the links from the initial complete graph, but weak at detecting "possible causes" through colliders (in part of because of our conservative approach as described above), and moderately effective at detecting "definite causes". However, our explorations revealed that these algorithms leave behind a fair amount of structure that is unexploited, e.g., we may know that a triplet of variables will turn to be either *a* -> *b* -> *c* or *a* <- *b* <- *c* or *a* <- *b* -> *c*, but without help in orienting one of the two links, the other link can't be oriented., so the algorithm simply produces *a* – *b* – *c*, undirected. However, the researcher may have very strong substantive grounds for asserting that *b* does not cause *a*¸ e.g., because *b* may be an event that chronologically preceded *a*, which would immediately suggest the orientation *a* -> *b* -> *c*. This suggests the need for incorporating reliable substantive knowledge into the algorithm, e.g., manually setting an assertion *a* -> *b*, and then reiterating the IC+ algorithm to automatically derive inferences such as *b* -> *c*. However, is obvious that injecting any assumption at all into the process is the first step down a slippery slope which can result in `discovering' a model concocted by the researcher, precisely the garbage-in garbage-out problem that our methodology seeks to avoid. The practical question then is how best to inject such substantive knowledge into the IC+ algorithm. The problem cannot be avoided because we need to construct a completely oriented causal graph in order to generate the structural equation model whose parameters will be estimated. Whether we orient the graph through automated discovery or through manual assertion, orientation must be complete.

The risks of injecting the researcher's favored theories into the model can be mitigated by implementing several principles in our methodology.

1. Automatic inference from empirical data must always override theoretical assumptions whenever possible. Therefore the algorithms implementing our methodology will ignore a theoretical assertion that orients a link whenever the link's directionality can be inferred by the IC+ algorithm from the given data (plus any previously accepted assertions, depending on the state of analysis). In fact, when an automated inference contradicts an assertion supplied by the modeler, it reliably indicates that we have either a significant insight, or an error in an earlier stage of analysis. One implication of this principle for implementation is that for every assertion that we inject into the analysis, the IC+ algorithm must be run to its conclusion, in order to maximally extract as much causal information as possible from the data plus assertion.

2. Assertions must be organized in terms of reliability or defensibility, and the most trustworthy assertions must be applied first. This sorting increases the likelihood that less reliable assertions lower down the list end up not being used at all, because the initial assertions may turn out to be sufficient for ensuring that IC+ orients the whole graph. This goes back to our earlier point that in practical use we found that there is a large amount of causal structure left unused by IC+, and a injecting a few theoretical assertions can trigger very effective detection of this residual causal structure. The better we are at triggering detection of causal

structure, the fewer the assertions we need. Using a handful of trustworthy theoretical assertions to automatically orient large parts of the graph substantially raises our confidence in the quality of the entire model, because a model is as weak as the weakest assertions that it contains.

One extension of this principle is to model our confidence in each assertion as a probability value, and utilize a Bayesian approach to how we measure our belief in the entire model depending on which assertions were applied. We did not formally develop such an approach; instead we used a much simpler heuristic implementation, and grouped assertions into qualitative confidence levels on an ordinal scale, e.g., "Safe", "Reasonable", and "Ambiguous" assertions, applying them in that order. The algorithms of course ignored assumptions that were not needed per Principle 1. In our particular of consumer purchasing, we typically found that we had grounds to label about half the assertions "safe", and most of the other half "reasonable", with only a small fraction truly ambiguous where we did not know how to orient the link.

Another implication of this principle is that it engenders an optimization problem that would be solved in the ideal implementation of our methodology. (We did not actually solve this problem, instead utilizing the above heuristic approach.) Given a partially oriented causal graph produced by applying the IC+ algorithm, what is the best link that should be manually oriented next, in order to trigger the maximal use of unexploited causal structure? Orienting one link may result in no new inferences, but orienting another link may trigger a whole sequence of new directional inferences. A look-ahead algorithm can conceivably be designed to score each undirected link on the number of new inferences it will trigger when oriented; the link with the highest score would be presented to the modeler for their input regarding directionality. The optimization problem gets complex when we consider sequences of assertions, since the order in which assertions are introduced affects the need for subsequent assertions; the minimal subset across all possible permutations of assertions is desired. If we factor in a Bayesian belief level for each assertion, the problem becomes very difficult---not from a theoretical perspective (since we need to just solve an optimization problem to find the assertion sequence with the greatest net belief score), but because we would need to acquire scores on all assertions in advance, imposing a tremendous load on the user, instead of only prompting for beliefs on a small subset of assertions on the remaining undirected links. The heuristic approach we implemented is a reasonable compromise in terms of reducing the modeler's effort, but leaves room for improvement in terms of obtaining the theoretically maximum confidence in the model for the least amount of effort.

3. Having the right vocabulary in which we can specify our assertions can substantially improve our confidence in the model. For example, we need the ability to assert that *A does-not-cause B* and *A dual-causes B*[26] in addition to *A*

[26] A "dual-cause" is a simultaneous pair of links in opposite directions, i.e., *A causes B* and *B causes A*. This occurs occasionally when we ignore time sequence in our models for simplicity. For example, a Visit to Retailer A may cause you to not Visit Retailer B (say because you found the right product at A), and vice

causes B, and *A possibly-causes B*. In practical use we discovered that the *does-not-cause* assertion is extremely powerful and used 95% of the time; *common-cause* and *dual-cause* assertions are used occasionally, and direct causal assertions *(A causes B)* are never used. The reason for the power of the *does-not-cause* assertion is that it directly translates chronological sequence (if *A temporally precedes B*, we generally cannot comfortably assert that *A causes B*¸ but we can with 100% confidence assert that *B does-not-cause A*), and in general directly captures a lot of reliable domain knowledge that is not temporal. For example, we may be unsure whether a salesperson's appearing *knowledgeable about the products* may have caused you to *feel the salesperson was trustworthy*, but it is unlikely that trustworthiness causes a perception of knowledgeability; the latter is much more likely to be unrelated factors such the training received by the salesperson. Because our confidence in *does-not-cause* assertions is generally much higher than in *causes* assertions, relying on the former increases our overall confidence that we have not injected indefensible theory into the model.

4. The impact of an assertion on the model should be made visible, in at least two ways. (i) Any alternative assertions that could have been applied to orient the same link in a different way, e.g., *A causes B* could have been plausibly replaced by or augmented with an *A has-common-cause-with B* assertion. (ii) The other links affected by this assertion, e.g., by automatically generating inferences through IC+, should be visible. This is important because injecting one incorrect theoretical assertion does not generally destroy the whole model; it only destroys part of the model, typically the part that is adjacent to the affected link. This is a consequence of the local nature of causal models; we may get one mechanism wrong, but other mechanisms elsewhere in the model are not affected since the intervening nodes isolate their effects on each other. Therefore there are great benefits to be obtained from deriving a good intuitive understanding of which parts of the model are robust and which are weak due to unreliable assertions. In particular, some critical inferences may turn out to be resilient to these errors, because changing the assertions may not change the conclusions, justifying robust conclusions even from a partially faulty model. When changing an ambiguous assertion does have an effect on the model, it is important to at least know the range of conclusions that can be drawn. Implementing this principle implies that full traceability must be maintained so that the researcher can begin to visualize and internalize the impact of an assertion.

These principles led us to design an interactive modeling methodology which embeds the fully-automated IC+ algorithm into surrounding interactive algorithms. These algorithms (a) identify a small set of undirected links for which the user is asked to provide substantive assertions; (b) apply the assertions in order of decreasing trust, (c) invoke the IC+ algorithm whenever possible, maximizing automated discovery of directionality, and (d) iterate suitably until all the entire graph has been oriented. All of this has been

versa. Information on the sequence in which the stores were visited could be used to resolve the circularity, but it is simpler to permit arrows both ways and rely on the SEM tools to estimate their respective effect sizes whenever possible---which, surprisingly, is most of the time.

summarized as “Phase 2: Causal discovery guided by substantive information” in the flowchart of our quantitative methodology (Figure 19).

4.1.2.4 Estimating the model

The third and final phase of our quantitative modeling methodology is specifying the functional form of the relationship between each variable and its direct causes, and then estimating function parameters. The simplest form, which we use by default, is a linear relationship: each variable is written as a weighted sum of its causes. There are several exceptions to this, notably when variables are binary or categorical, and when quasi-deterministic functional relationships govern the variables. Our methods for handling these and other nonlinearities are described in Section 4.3.6. In general, the set of equations relating each variable to its parents, plus the residual covariance relationships (assumption that a common-cause may exist, or that it doesn’t) between every pair of variables constitutes a structural equation model (SEM) that can be automatically generated from the causal graph obtained in the previous phase. Our causal modeling toolkit in R exports the equivalent SEM in Mplus input language along with an accompanying dataset that can be immediately imported into the Mplus software and executed. In addition to obtaining individual link parameters, our code is programmed with an option for obtaining the total effect of each variable on the designated primary outcome (*Sales*), which is particularly useful for understanding the effect of an intervention on each variable.

We used the maximum likelihood estimator wherever it was supported in Mplus Version 3, i.e., MLE was used for linear models, and Generalized Least Squares was used when categorical outcome variables (and therefore the probit function) were present. In general we did not have convergence problems; the exceptions were when we had a number of *dual-causes* (feedback loops) or when we had categorical variables on both the left and right hand sides of an equation.[27] Minor manual changes to the model (see Section 4.1.2.5) usually fixed the first of these problems; however we sometimes had to compromise and drop less-important categorical variables to work around the latter problem.

The Mplus output also provides standard measures about the goodness of model fit, in addition to providing a lot of useful data on the reliability of each link parameter. There are numerous measures of model fit in the SEM literature (e.g., [Bollen 1989, Garson 2006]). We relied on several of the most commonly used ones, viz., Model Chi-square not less than 0.05, Root mean square of approximation (RMSEA) less than 0.05, Standardized root mean residuals (SRMR) less than 0.01, Comparative fit index (CFI) and Tucker-Lewis fit index (TLI) both greater than 0.9. While of these measures only look for mismatch between our model specification and the given empirical data, others also measure model parsimony, e.g., penalize having too many links in the model specification, which would make it easier to fit any dataset. In addition to these

[27] The latter appears to be an Mplus bug; newer versions of the software have improved on the convergence issues.

measures, we also relied on the R-square of each outcome variable, and the modification indices generated by Mplus to get a sense of how well the model fit the data; the former measures the degree to which we have succeeded in explaining the causes of a given variable, and the latter identifies links that might have been mis-specified in our model. In passing note that none of these measures "proves" that we have the right model, since there are always multiple models that will fit a given dataset equally well. What these tests prove is that we don't have a wrong model that fails to fit the data, and therefore raises our confidence in the model chosen.

As the model size and sample size increases, it becomes much harder to correctly specify a model that passes all these tests, e.g., does not trigger a modification index on any of the thousands of links tested. The increased likelihood of falsification correspondingly increases our confidence when we do succeed in finding a model that fits. Therefore it was with some surprise that we observed that SEMs generated via our causal modeling methodology (the previous two phases based on IC+) succeeded in passing all these tests of goodness right on the first try, without manual tweaks of the kind described in the next section. From this experience we infer that causal discovery algorithms such as IC+ are very effective at avoiding poor models and narrowing down the model space to just those models that will fit the data very well. This also gives us good reason to be optimistic about the practical effectiveness of causal discovery algorithms in general, despite the criticisms of Freedman and others, although we will remain somewhat cautious until our algorithms have been tested on a wide range of datasets.

4.1.2.5 Modifying the model

While it is remarkable that the SEM autogenerated by the causal modeling algorithms often turned out to have excellent fit on the very first estimation run through Mplus, sometimes the model fit turned out to be unsatisfactory, and we were required to make modifications, regenerate a new SEM, and test model fit, iterating until the fit was satisfactory. In traditional SEM methodology, modifying model structure is largely a black art, with vague guidance provided by modification indices, but largely relying on the researcher's ability to make a better guess at the true model structure. Inevitably, the scientific rigor of the modeling process deteriorates (see detailed discussion in Section 4.1.2.7).

In contrast, in our approach we experimentally discovered, to our pleasant surprise, that poor model fit was usually due to a particular class of modeling errors that could be systematically pinpointed and corrected using diagnostics available from the estimation software and the causal modeling algorithms. In other words, the guesswork has been mostly taken out. This substantially raises the value of the modeling methodology because of the increased scientific rigor.

The symptoms that trigger a diagnosis of these errors are the familiar modification indices. While traditionally, these indices do not imply a specific type of model correction, in our approach, a high index on a link between two variables is usually diagnostic of an unwarranted independence assertion having been set between the two variables (i.e., an incorrect link deletion). The error is usually generated during the very first step of causal discovery algorithms, when the algorithm decides that the partial correlation between the two variables is insignificant conditional on some set of control variables. Our Causal Modeling Workbench provides a listing of what triggered the independence assertion. A visual examination of this trace usually reveals that the set of control variables which resulted in a near-zero partial correlation comprises variables that we would normally treat as causal effects of the two given variables (Figure 22). In other words, the zeroing of the partial correlation was a spurious result of the cancellation of two effects.

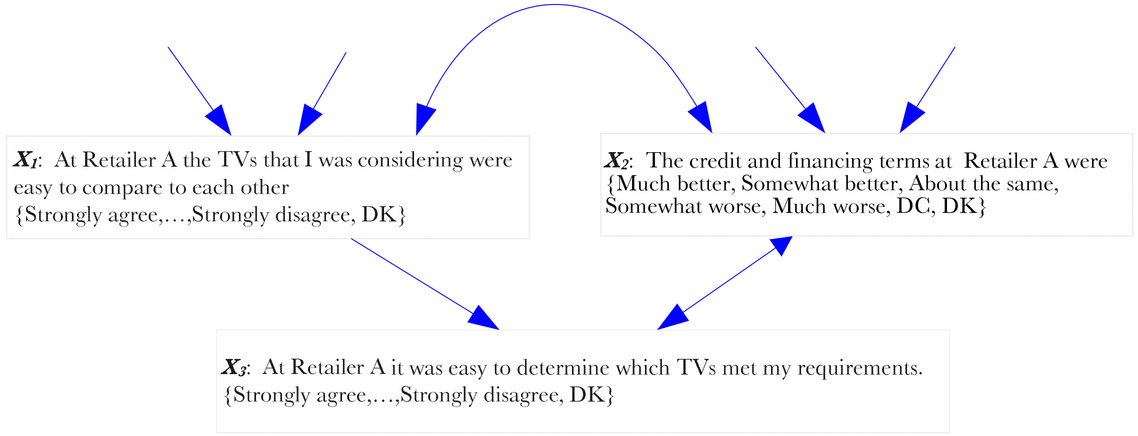

X_1: At Retailer A the TVs that I was considering were easy to compare to each other
{Strongly agree,…,Strongly disagree, DK}
X_2: The credit and financing terms at Retailer A were
{Much better, Somewhat better, About the same, Somewhat worse, Much worse, DC, DK}
X_3: At Retailer A it was easy to determine which TVs met my requirements.
{Strongly agree,…,Strongly disagree, DK}

Figure 22 Spurious cancellation: conditioning on an effect

X_1 should not be independent of X_2 conditional on X_3 since there is a collider at X_3. However, conditioning on X_3 induces a negative correlation between X_1 and X_2, which can cancel out against the positive correlation from the common-cause between X_1 and X_2.

This is a manifestation of the well known risk of violating the stability assumption (aka the faithfulness assumption) [Steel 2004]. In a model as small as 35 variables, there are over a thousand possible links and millions of conditional independence tests to be performed, which results in a nontrivial risk of a spurious cancellation. The larger the model, the greater the risk that some spurious conditional independence has caused an incorrect link deletion. The good news is that it is easy to spot and correct these deletions using the diagnostics from the tools. Once the link is restored (by setting a Dependency assertion using our normal assertion-injection procedures in Step 2), the causal modeling algorithms run their usual course through all the steps and generate a fresh SEM.[28] Upon estimation, the model displays good fit, presuming all instances of this type of error have been fixed.

It is worth noting that the modification indices were not used for any purpose other than identifying which two variables had the spurious link deletion. Further, no corrections were made to the model structure other than restoring that link, and sometimes adding a single substantive assertion. This minimalism adds to our confidence in the ability of these tools to discover the right causal structure.

[28] Sometimes the restored link does not get its directionality automatically detected, and the researcher may have to add another substantive assertion following usual procedure.

4.1.2.6 Managing model complexity

Given the large number of variables in our study, the computational performance of our structure detection algorithms became an important practical consideration, and we used heuristics to manage the complexity. Causal variables with minuscule correlations with the outcome were dropped when we had no reason to suspect nonmonotonic relationships or perfect cancellations (violations of the stability assumption [Pearl 2000]). We also limited tests of conditional independence to a maximum of three control variables, based on experimental observation that adding more control variables substantially increased computational space requirements without significantly changing the resulting structures. It's possible that a more efficient algorithmic implementation would avoid the use of these heuristics,[29] but they appeared to work quite well in our experience.

We also experimented with an alternative `bottom-up' or `modular' approach to model construction. First we built a series of local submodels (e.g., a submodel that comprised all the Price/Promotion variables and their interrelationships, another submodel for the Product variables, and so on), and then assembled all the submodels into a full model after dropping variables that were `upstream' in the model, i.e., variables that would be conditionally independent of the outcome given other intermediary variables in the submodel. Although this technique successfully reduced model complexity without the use of heuristics, we found this approach relatively unproductive in the end. The primary difficulty is that the upstream variables often have greater *total effect* on the outcome than their downstream intermediaries, often because the upstream variables affect the outcome through many paths passing through other submodels. Further, structural relationships between the variables in a submodel and the outcome may be altered by the inclusion of variables from other submodels in the full model. These properties make it nearly impossible to modularize the whole model into minimally-connected submodels *a priori* i.e., without having first built the full model, which is a chicken-and-egg situation. Therefore our present recommendation is to use the above heuristics to manage model complexity.

4.1.2.7 Additional benefits from our methodology

The semi-automated dependency structure identification step at the beginning of the methodology in effect replaces the manual modification of structure that is done at the end of the analysis process in traditional structural equation modeling. In traditional SEM, the researcher begins with a hypothesized causal structure and model diagram, specifying for each endogenous variable its hypothesized causes (and unexplained-covariance links). After estimating model parameters and determining goodness of fit, the researcher selectively drops or adds links between variables, guided respectively by low parameter values and high modification indices. This is a highly error-prone process, as has been documented by several researchers [Hutchinson 1998, MacCallum et al. 1992]. A structural decision such as retention or elimination of a link based on link

[29] Note however that there is another penalty for adding more tests of conditional independence: increased risk of statistical error, marking a link independent (conditional on some combination of variables) when it really isn't. If there isn't much to be gained in terms of structural change in the causal graph, the added risk may not be worthwhile. Resolving this tradeoff is a nice optimization problem.

strength is sensitive to the researcher's choice of which other variables are treated as covariates; further, each such structural decision itself affects parameter estimates in the models that are subsequently explored. In practice, this results in widely varying models being produced by different researchers, depending on the sequence and particulars of the modification process used by each individual researcher to generate a series of model improvements. This is a weak spot in the traditional SEM methodology, which substantially weakens the scientific rigor of the entire analysis process, and reduces the value of SEM as a modeling tool. There is a need to reduce the manual element in this process, since in general it is impractical to manually explore the entire space of possible models for a nontrivial problem, and this difficulty is what induces researchers to make their "best guess" at a model and end up "verifying" that particular model. Since the SEM method is designed to test a model and reject poor models but does not actually select among the remaining thousands of well-fitting models, it leaves researchers with ample ambiguity during model modification to move towards whatever theories are favored, and does not highlight contradictory theories that fit the data equally well.

The solution is a methodology that (a) explicitly but compactly visualizes, and supports examination of, the space of well-fitting models that remain after making any set of assumptions, so that it is evident at a glance what alternative theories are at play, and (b) reduces the manual elements in model exploration, by increasing automation of structure discovery, and by giving the researcher more disciplined controls (e.g., control of `risk' or error probabilities) in place of full control over every link in the model.

Compact model exploration results from the use of undirected links as well as directed links; the earlier stages of our modeling methodology use undirected links which implicitly represent all possible models in which these links are replaced by directed links. As assumptions are introduced to limit directionality on the undirected links, it becomes progressively clear where model structure is precisely defined, and where model structure remains ambiguous supporting alternative theories.

Reduction of the manual element occurs as a byproduct of the semi-automated discovery process we have used: in our analyses, we regularly found that the auto-generated SEM had excellent fit either with no manual modifications at all, or with a few modifications (unless the risk controls were substantially off their typical values, or there were pathologies in the data as mentioned earlier). So we did not have to mess around much with model modifications, and when we did need to improve the model we could do so in a disciplined manner by modifying an explicit assumption in a way that was guided by the data (see Section 4.1.2.5). In effect, the independence structure discovery algorithms implicitly investigate the space of many possible causal structures that would traditionally have to be explicitly generated and tested manually before deciding to eliminate a link in the model. When these algorithms drop a link, in effect they eliminate all possible causal structures that would have contained arrows along this link. Likewise, the second step in which substantive assertions are introduced in order of defensibility also acts to eliminate poor models. Using this framework of progressive elimination in place of the `generate-and-test' framework[30] helps in two ways: it keeps

[30] For further insight, consider the relationships between the Branch-and-Bound and Generate-and-Test paradigms used in Operations Research and Artificial Intelligence respectively.

the modeler always within the space of valid models, and keeps all the alternative theories in full view.

4.1.3 Measuring the "importance" of a variable

After the causal model has been built, the next step is typically that of using the model to infer the best possible business intervention(s) that would maximize the outcome, in our case sales. Decision makers typically ask, "What is the most important factor that drives sales?" "What are the top ten critical factors?" Answering this is a lot trickier than it sounds.

Part of the problem is that a concept as common as "importance" is mathematically ill-defined. Is the importance of a variable the "relative weight" with which it loads on the outcome variable, i.e., the regression coefficient, or path coefficient in a structural equation model? Suppose we compare two variables as diverse as *Spoke to a salesperson* and *Felt I got a good deal*, even though the causal model will tell us how much a *unit change* in each variable will increase *Sales*, how do we make an `apples-to-apples' comparison when the scales of the two variables are so different? Indeed what is a "unit change" on a binary scale such as the first variable; e.g., is it a 1% increase in the frequency of talking to a salesperson? And on the second variable which is 5-point Likert scale, a 1-point increase does not even have consistent meaning at different parts of the scale because the scale is ordinal. In any case, why would we compare a 1% intervention on one variable to 1-point intervention on another scale? If we choose to compare a 10% intervention on the first scale to a 1-point intervention on the second scale, the resulting conclusions about which variable is important could be reversed, since the first variable would now produce a greater increase in sales. Clearly, we need a meaningful way to compare an intervention on apples to an intervention on oranges, in order to determine which of the two is "more important".

Most studies commonly adopt one of two approaches to measure the importance of a variable. The first simply asks people to judge "importance" directly. For example "was *Price* more important than *Service* in influencing your decision to buy from Retailer A?" Alternatively, respondents are asked to rank the variables "in order of importance", or to rate the importance of each variable on a 5-point scale from "Very important" to "Unimportant". All of these approaches based on direct judgment have severe deficiencies from a causal perspective. For example, suppose we consider making a $10 improvement in *Price*, and suppose that *Price* was given the highest rating on the 5-point importance scale, what will be resulting increase in *Sales* which is measured in dollars? Do we multiply the value of *Price* by the value of *Importance of Price*? And if we similarly consider a 1-point (Likert-scaled) improvement in *Service*, do we multiply the *Service* variable by its importance rating in order to compute the resulting increase in *Sales*? Not only is there still a problem comparing disparate scales, there is a more fundamental problem: do these judgments of importance truly have a corresponding causal effect on sales? For example, it is possible for *Price* to be judged to be "extremely important" yet to have little effect on *Sales*, e.g., because all retailers offer the same price.[31] The problem is that the concept of `psychological importance' is not the same as the concept of 'causal effect'. Even when a variable feels psychologically "very important" to a shopper, a change in the variable may not actually induce a change in the outcome. Shoppers in our study told us that *TV size* was very important (in order to fit their TV cabinet), yet, since all major retailers offer most sizes of TVs, this variable had only a small effect on the choice of retailer. From a modeling perspective, *TV size* is just the price of entry; if you didn't have any TVs in the right sizes you'd lose the customer and therefore the variable is "important". Yet, given the fact that all sizes are normally available, the variable will not actually affect the outcome, and no intervention on the variable is needed; hence the variable is unimportant. From a business decision maker's perspective, it is of little use that a variable is psychologically important to the consumer if an intervention on the variable actually produces no improvement on sales.

The difficulty can be better understood from a modeling perspective if think about the semantics of links in a causal diagram. A link denotes the effect of one variable on another, and in linear models, the link weight represents the strength of an effect. When we ask people to judge the importance of a causal variable, we are in effect asking them to estimate the weight on the link from the given variable to the outcome. On the other hand, when we solve regressions and SEM models, we infer the link weights via the estimation process. In other words, the effect of a variable is inferred from the covariation between the two variables. If a change in the given variable does not correlate well with a change in the outcome variable, the link weight will go down. It is no surprise that a shopper's judgment about link strength would be at odds with a statistically derived estimate of link strength that reflects actual covariation between the variables. In short, it is a modeling error to elicit direct judgments about importance and then try to draw causal inferences from those psychological judgments.[32]

[31] This is common in many retail sectors including electronics: the product manufacturers enforce a maximum advertised price (MAP) regimen on retailers to prevent price wars.

[32] Unless the causal mechanisms that relate the psychological judgments to actual behavior are known. If that were the case, the mechanisms would be modeled explicitly, and the importance judgment would be a *variable* like any other, not a *link strength*. The mechanisms that relate the judgment to the behavioral outcome would be represented by links, with link strengths that are again estimated via statistical methods, not judged directly by respondents.

There are a few such circumstances when a multiplicative model utilizing direct judgments may be appropriate. For example, when respondents are asked how they felt about the *Sound quality* of a TV, they are also given the option of answering "Didn't care" instead of rating the sound quality on a 5-point Likert scale. Intuitively, if a shopper did not care about sound quality, any rating of sound quality should not affect other variables such as the purchase decision. In other words, any causal links from *Sound quality* to downstream variables should be weakened or removed.

This is handled in a natural way when we recognize that *Cared about sound quality* is a quasi-deterministic switch variable that controls the effect of *Sound quality* on other variables. In our model, we did not elicit a rating on *Sound quality* from respondents if they did not care about the attribute; in other words, *Sound quality* is quasi-deterministically missing (Section 4.1.4) based on *Cared about sound quality*. The ensuing switched-product analysis method shows how a direct judgment of importance such as *Cared about sound quality* can be used to indirectly weaken links in the model while carefully maintaining correct causal semantics. From a modeling perspective, the justification to using direct judgment here is the knowledge of the causal mechanisms by which psychological constructs influence behavior. Note in particular that respondents are not asked to judge the causal effect of one variable on another; we just examine whether

The second approach commonly used to measure the importance of a variable is typically recommended by statisticians who construct regression models including SEM: standardize the regression coefficients and declare the variables with higher standardized coefficients as being more important than the others. In support of this approach, it is commonly pointed out that the coefficients do reflect variance-explained, i.e., an explanatory variable with a higher coefficient accounts for a greater proportion of the variation in the dependent variable, and thus better "explains" the outcome.

We believe this approach is also erroneous, for several reasons. First, the coefficients in a regression, or path coefficients in an SEM represent the *direct* effect, not the *total* effect. In other words, the coefficient represents the effect that a given variable *X* would have on the outcome *Y* if all other variables *Z* in the model were *kept constant*. This is usually a ridiculous intervention to consider---why or how would we ever keep the rest of the world unchanged when we try to intervene on one variable? When a business decision maker asks "what is the effect of a change in *X*?" the intended common-sense translation is "I'm going to change *X*, and let the rest of world *Z* work as it normally would[33]; what is the effect on *Y*?" In causal modeling language, a decision maker's need to understand the "effect" of a variable should be translated as measuring "total causal effect", i.e., $P(Y|\mathrm{do}(X))$. While this is strictly speaking a probability distribution that shows how the world will turn out if we make the intervention on *X*, we can simplify this in linear SEM as being represented by the total effect coefficient, i.e., the sum of all path coefficients from a given explanatory variable to the outcome variable.

Using total effects instead of direct effects still does not resolve the issues with utilizing standardized coefficients as measures of importance. The second problem with this approach has to do with considering the *size* of an intervention. When using standardized coefficients, common practice is to use 1 standard-deviation as the unit of intervention for all variables. Our view is that is a meaningless exercise. Why would a business decision maker consider a 1-standard-deviation intervention on *Spoke to a salesperson* equivalent to a 1-standard-deviation intervention on *Felt I got a good deal*? Recall our earlier discussion where we argued that considering a 10% increase in frequency of *Spoke to a salesperson* is no more special than considering a 1% increase; why does the variance of the variable define an intervention size that makes any more business sense?

That the standard deviation is irrelevant becomes more evident if we consider how hard it might be to make an intervention. For example, if it costs millions of dollars in advertising to change consumers perceptions of how good a deal they are getting---by 1 point on the 5-point Likert scale---but costs little to train salespeople to increase frequency of contact with customers---by 10%, then it might make more business sense to invest in training salespeople even if the standardized coefficient of this variable is lower than that of the other variable. Clearly, standardized coefficients do not reflect the *cost* or *difficulty* of making an intervention. The fact that a higher coefficient reflects greater explanatory power does not mean that the corresponding variable is a better

they even cared about the product attribute and draw causal inferences from that. For more on this, see Section 4.1.4.7.

[33] For example *X* may affect *Z* and thus influence *Y* indirectly through changes in *Z* in addition to any direct effect that *X* may have on *Y*.

investment. In order to enable comparison of coefficients, critical information about the scales of the variables had to be discarded. We no longer know the meaning of a 1-unit change in a variable (other than that it represents 1 standard deviation); therefore we can no longer reason about how easy or difficult the change will be, how much it will cost, or in short, how practical the intervention is. In an effort to compare apples to oranges by declaring them all as fruit and discarding more specific information, standardized coefficients have lost the basis for making decisions.

In our view the only way to compare variables that are on totally different scales is to find some standardized *basis* for comparing *interventions*, not necessarily to standardize the variables themselves.[34] To our minds, currently only one such criterion stands out: the `return on intervention', i.e., the gain in sales (measured in dollars) divided by the cost of making an intervention (also measured in dollars). Thus, if we wish to compare the "importance" of *sending advertising flyers* against *giving the perception of a good deal*, we need to first estimate the dollar cost of sending one extra unit of flyers, the effect on sales of a 1-unit increase in flyers, and thus the return on the intervention. Likewise, we need to do the same for a 1-unit improvement in perception. When we compare the two numbers, we are in effect comparing apples to apples: how much a dollar invested in each variable yields in sales dollars. It may turn out that the variable with a high total effect coefficient is enormously expensive, and hence its return on investment will be small. On the other hand, a variable with a tiny effect coefficient may also be extremely cheap and therefore a good investment choice, which business people understand quite well as "going after the low-hanging fruit."[35]

As full disclosure it is worth mentioning that in this particular study we used standardized total effects as the basis for identifying critical factors; see Section 4.3.4 for details on the difficulties we encountered. However we would avoid that method that in future studies, and stand by our above recommendations about the best way to measure the importance of a variable.

4.1.4 Quasi-deterministic models

Section 3.2.3 introduced the concept of quasi-deterministic (QD) functions, the part-deterministic part-stochastic functional relationship between two or more variables, and described their importance as a new modeling construct, by analogy with how logit transformations provide better modeling capabilities than linear models. We now develop the analytical tools for modeling quasi-deterministic relationships.

4.1.4.1 Definitions and some properties

Before providing a formal definition of QD functions, it is useful to recapitulate the more basic definition of a causal model [Pearl 2000].

[34] If interventions on multiple explanatory variables have additive effects on the outcome, we can standardize the variables themselves by weighting them by their respective costs of intervention. However this is not generally true and we therefore have to consider the total cost of simultaneously intervening on a group of variables versus the total increase in sales that such an intervention would generate.

[35] Pun intended.

A causal model is defined via a combination of

- Structural specifications: a graph indicating which variables X are parents of (i.e., directly affect) a variable Y
- Functional specifications: the precise value of Y is determined by the function $f(.)$, thus $y = f(x, u)$ where U is an error variable that captures all causes of Y other than those represented by X. U is unobserved, i.e., it is always a latent variable, and is governed by some probability distribution $P(U)$.

Note that the equation $y = f(x, u)$ is a deterministic relationship, i.e., Y is assumed to be always determined with complete certainty by X and U. However, since the value of U is always unknown by definition, it follows that the value of Y is uncertain, and is described by a probability distribution $P(Y)$ which is determined by both $P(X, U)$ and the functional specification $f(x, u)$.

Using the above notation, we define a quasi-deterministic function as follows:

Equation 1

$$Y = \begin{cases} f_1(X) & \text{if } X \in \{x_0, ..., x_n\} \\ f_2(X, U) & \text{if } X \notin \{x_0, ..., x_n\} \end{cases}$$

The upper expression is the `deterministic' part, and the lower one the traditional stochastic part. In the upper expression, U is missing, thus indicating that the value of Y is determined with complete certainty by X, as long as the value of X is within a particular subdomain $\{x_0, ..., x_n\}$. Note that X may vary according to a probability distribution $P(X)$ thereby inducing a distribution $P(Y)$. However, the value of Y is precisely determined by X when X is within the given subdomain, e.g., $P(Y=y \mid X=) = 1$ if $y = f_1(x_0)$ and 0 otherwise. When the value of X is outside the given subdomain, the error term u is present, and thus the lower expression $f_2(X, U)$ is a standard functional specification as described earlier.

Note that one consequence of this definition is that there are empty cells in the joint probability distribution of X and Y. Let us define the *deterministic range DR(Y)* as the values $\{y_0, \ldots, y_m\} = f_1 : X \in \{x_0, \ldots, x_n\}$. Then Y does not take the values $\{y_{m+1}, \ldots, y_k\}$ outside that range unless X is outside its deterministic subdomain, i.e., $P(X \in \{x_0, ..., x_n\}, Y \in \{y_{m+1}, ... y_k\}) = 0$, due to the deterministic part of the QD relationship. Recall our example that the probability of not shopping at a retailer but purchasing from that retailer is nil; the two assertions are in fact logically inconsistent and that is reflected in the empty cells in real-world datasets. This consequence is worth noting because many important statistical definitions and analyses, including some of the properties of conditional independence and the uniqueness of Bayesian/causal graphs, require strictly nonzero probability distributions.

A special case of QD functions is that of quasi-deterministically-missing (QDM) functions: Y is *undefined* when $X \in \{x_0, \ldots, x_n\}$, in other words, $DR(Y) = \{\}$. Thus the `deterministic' upper component says that Y is (deterministically) missing whenever X is within the given subdomain. Recall our example that the *salesperson's friendliness* is meaningless when the shopper did not *talk with a salesperson*, e.g., if the shopper only visited the retailer’s Website. The value of the *Friendliness* variable is missing for all shoppers for whom the value of the *Talked* variable is *No.* Moreover, the `missingness’ of *Friendliness* is deterministically dependent on *Talked*, unlike the randomly missing values that are common in surveys because respondents occasionally skip a question.

The concept of QDM functions is not so important when we need to just predict *Y*; we could just use the lower half, i.e., the stochastic f_2 regression since it is meaningless to predict the value of *Y* when *Y* is inapplicable in the upper half. However QDM functions are very important when we need to predict some downstream consequence of *Y*, e.g., $Z = f(Y, X, u_2)$. In this case *Y* has an effect on *Z* whenever *Y* is defined, and has no effect when *Y* is missing; *X* an effect in both situations.

The “switching” metaphor is particularly useful in intuitively grasping the functioning of QDM functions. When *X* is a binary variable *X* it can be viewed as a switch that controls the functioning of other variables *Y*. *X* itself has a stochastic effect on *Z*. *X* “switches off” *Y* when *X* =0; as a consequence, *Y* has no effect on Z when *X=0.* When *X* is switched on, it enables *Y* to influence *Z* in the usual stochastic manner.

Note that quasi-deterministic models are quite different from the sample-selection models [Bierens 2002] censored data models, and truncated models that are common in econometrics. In sample-selection models, although the dependent variable *Y* has missing values (for some value of an independent variable *X*), the dependent variable is merely unobserved, not nonexistent. In other words, there exists a latent variable *Y** whose values are never missing, and *Y* is the observed component of *Y**. The task in those problems is to estimate *Y** and we can statistically impute or otherwise estimate the missing values. In quasi-deterministic models, there is no separate latent variable, because the values of *Y* are not merely unobserved, they do not exist logically and deterministically, because of *X*. QD models are different from censored and truncated models for the same reason. An additional difference from censored and truncated models is that in those models *Y* is missing according to some criteria based on *Y* itself; the missingness is not based on some independent variable *X.* Quasi-deterministic models are also different from two-part models [Schafer and Olsen 1999]; see footnote 37.

4.1.4.2 Missing-value methods

Quasi-deterministically missing (QDM) data appears to pose a serious problem during statistical analysis, for several reasons. First, huge amounts of data are `missing’, unlike the small percentage of missing responses that is typically present in surveys due to respondent errors. For example, since only 32% of the survey respondents had visited Sears, it followed that 68% of all cases had missing data on the in-store variables, because it was meaningless to ask shoppers what happened inside the store if they hadn’t visited it. Of the 32% who visited the store, 30% did not seriously consider even a single TV at the store, and therefore those shoppers did not provide data on the TVs they considered at that retailer. In other words, the data is relatively sparse and missing values cannot be viewed as a small `error’ in the analysis. Given our intuitive understanding of the causal mechanisms behind *why* the data is missing, it should be obvious that the `missingness’ is not a form of `error’ in the first place.

The last point further illustrates the second difficulty with analyzing the missing data. The statistical literature on the subject [Allison 2002, Schafer and Graham 2002] is limited to stochastically missing data, presuming that whatever caused the data to be missing was a probabilistic phenomenon. Thus the various forms of missingness are defined as MAR (missing at random), MCAR (missing completely at random), and so on. These definitions are used as the basis for developing missing-value analysis techniques such as listwise and pairwise deletion, dummy variable adjustment, imputation, maximum likelihood estimation, and so on. However, our data is not missing "at random" in any sense of that term; it's missing deterministically. If a shopper *did not shop* at a retailer, there is no data on *helpfulness of salespeople* at that retailer, with absolute certainty. If we do listwise or pairwise deletion, we end up throwing out all shoppers who did not visit the retailer (68% of the cases as we mentioned earlier, based on Sears alone; retaining only shoppers who visited all retailers leaves almost no data). Apart from losing data, the analysis would obviously be extremely biased if we only retained shoppers who visited a retailer. If we tried to impute the value of the missing data (e.g., substitute a mean value, or try to predict it based on other variables), we end up in a meaningless exercise because it is fundamentally meaningless to assess the *helpfulness of salespeople* for a shopper who doesn't visit the retailer. Thus any statistical technique that fails to recognize the deterministic nature of the missingness is inapplicable.

However, one of the techniques used for MAR and MCAR data, namely dummy-variable adjustment, provides the seed of a method for handling QDM data. In the dummy-variable technique, an indicator-variable *D* is created such that *D*=0 when *X* has a missing value, and *D*=1 when *X* has data. Then *X* itself is replaced by another variable *X** such that *X** = *X* when data is available, and *X*=*c* when data is missing, where *c* is an arbitrary constant. Analysis methods are designed to be invariant with respect to the choice of *c*. Both *D* and *X** participate in regression models. Jones [1996] proved that dummy-indicator methods are an unacceptable means for treating missing values because they introduce large biases. However, the analysis developed by Jones to compute the extent of the bias is also useful for demonstrating the bias-free nature of the approach that we develop below for the treatment of QDM data.[36]

The following sections will continue our analysis of quasi-deterministic relationships; for a treatment of traditional missing values see Section 4.1.8.

4.1.4.3 Treatment of QDM data for multiple regression

For simplicity of analysis, let us formulate a causal relationship involving quasi-deterministically missing data as follows:

Equation 2

$$Y = \begin{cases} b_0^{'} + b_1^{'} X_1 + b_2^{'} X_2 + e^{'} & \text{if} \quad S_1 = 1 \\ b_0^{''} + \qquad\quad b_2^{''} X_2 + e^{''} & \text{if} \quad S_1 = 0 \end{cases}$$

[36] The Jones paper, and its use in confirming the bias-free nature of our proposed approach, was brought to our attention by Ken Bollen.

where *Y* is an independent variable, and $S_{1,}$ X_1, and X_2 are explanatory variables, and X_1 is always missing whenever $S_1 = 0$. While S_1 is a regular variable no different from X_1, X_2, or *Y*, we have given it a distinct notation here to help the reader recognize its role as a "switch" variable, which turns off X_1. In particular, note that S_1 is not a "dummy" variable created solely for the purpose of analysis to indicate missingness as in the missing-indicator methods. For example, S_1 may be *Visited the store*, X_1 may be *Friendliness of the salespeople*, X_2 *Saw advertising*, and *Y Purchased*. Since data on salespeople cannot be available without a visit, X_1 is deterministically missing as controlled by S_1, but X_2 and *Y* always exist. For simplicity, this formulation assumes that the switch is a binary variable (*Visited the store = True* or *False*, numerically coded 1-0), but a slightly more complex formulation permits the treatment of continuous-scaled switch variables as well.

Clearly information about the missingness of X_1 is already present in the regular variable that caused X_1 to go missing, namely $S_{1,}$ so we do not need to create any "dummy" variables. Also, we can populate the missing data cells in X_1 with any number of arbitrary values, but our analysis will not be affected by that. In practice, we do substitute some numeric code, because most statistical software packages are not programmed to deal with empty cells (e.g., they are unable to infer that the product of a missing-value and 0 must be 0) and by default would delete the case. Supplying an arbitrary non-missing value is an expedient workaround for computational purposes, but does not affect the theory.

The above equation actually describes two regressions, corresponding to a partitioning of the dataset into two segments (sets of cases). The first regression, corresponding to $S_1 = 1$, by definition comprises the set of all shoppers who visited the store, and involves both X_1 and X_2. The second regression comprises the shoppers who did not visit the store, and hence only X_2 is present. Although intercepts, coefficients for X_2, and error terms and present in both halves, they are different in each half, in general; hence the use of primes and double-primes in the notation. Actually estimating parameters via two regressions and combining the resulting "submodels" via the above equation produces a perfectly meaningful and correct "full-model", and would be an acceptable way to handle simple situations involving QDM data.[37] Unfortunately, when many such QDM relationships are involved as described in Section 3.2.3, the number of such submodels explodes, and manual assembly of submodels is impossible. Nor is it possible to apply structural equation modeling techniques, solving simultaneous regressions on large models, using such two-part models.

The upper and lower halves can be combined into a single equation as follows:

[37] Note that this two-regression model is different from the "two-part models" common in econometrics, e.g., [Duan et al. 1983, Schafer and Olsen 1999, Zhou et al 2004]. In the latter, the two parts are the binary regression for the switch *S* and linear regression for the dependent variable *Y*; there is only one regression for *Y*. We have two regressions for *Y* corresponding the two values of the switch *S*. Of course we have a binary regression for *S* as well, but we don't call it out in our discussion of QD models here because it is merely a special case: every intermediate variable in a structural equation model has its own regression equation, and the switch variable *S* is just another intermediate variable to be explained by its own causal parents.

Equation 3

$$
\begin{aligned}
Y &= \left(b_0' + b_1' X_1 + b_2' X_2 + e'\right) S_1 + \left(b_0'' + b_2'' X_2 + e''\right)\left(1 - S_1\right) \\
&= b_0'' + \left(b_0' - b_0'' + e' - e''\right) S_1 + b_1' X_1 S_1 + b_2'' X_2 + \left(b_2' - b_2''\right) X_2 S_1 + e''
\end{aligned}
$$

This is a single regression in which S_1 and X_2 participate as expected, X_1 only appears as a product term (interaction term) with S_1, and X_2 appears both by itself and as in interaction with S_1. This is intuitively meaningful, and several interesting observations can be drawn:

- Since X_1 only appears as a product with S_1 and since S_1 is zero whenever X_1 is missing, it follows that any arbitrary numeric codes that have been substituted in the missing data cells of X_1 will not affect the regression.
- The coefficient of X_1S_1, namely b_1*'*, is actually the coefficient of X_1 in the *upper submodel* of the original equation. Thus the coefficient of X_1S_1 is not really interpreted as an "interaction effect" as with typical regressions, but actually interpreted as the main effect of X_1 whenever X_1 applies, i.e., whenever the switch S_1 is "on". In our example, b_1*'* is the effect on purchase of the friendliness of the salespeople for those shoppers who *do* visit the store. This is perfectly meaningful since friendliness cannot have an effect on shoppers who don't visit anyway. If we had built a submodel comprising just the shoppers who visited the store, we would have obtained the same coefficient for X_1; however we are now reading off the coefficient from a full-dataset regression by looking at X_1S_1 instead of X_1. This is an important general principle that will be useful later in developing the treatment of structural equation models of QDM data:

  > Switched-product rule: X_1 can be replaced by X_1S_1 in many analyses, provided the resulting coefficients are interpreted and handled carefully. In particular, the mean-normalized switch-product (X_1 – E[X_1])S_1 will be introduced in Section 4.1.4.6 as the basis for SEM analysis.

- The coefficient of X_2 in the above equation namely b_2*"*, is actually the coefficient of X_2 in the *lower submodel* of the original equation, and is thus interpreted as the effect of X_2 when the switch is "off". However, X_2 also has an effect when the switch is on. This is implicitly reflected in the other appearance of X_2 in the above equation, namely the coefficient of X_2S_1. That coefficient, which is equal to (b_2*'* – b_2*"*) is interpreted as the differential effect of X_2 between the two submodels. In our example, *Advertising* may have some effect on shoppers who visited the store, and a different effect on shoppers who did not visit the store. If the effect were the same on both groups of shoppers, the coefficient of X_2S_1 would be zero. This observation is very useful for simplifying the model:

  > If we have *a priori* substantive or theoretical grounds for assuming that there should be no difference between the two groups of shoppers (or more generally, there should be no difference in how one variable X_2 affects another variable Y whether a switch is on or off), then we can make the simplifying assumption that we can drop the product term X_2S_2 from the regression. This results in a simplified regression for QDM data,

$$Y = b_0 + b_1 X_1 S_1 + b_2 X_2 + b_3 S_1 + e$$

which appears identical to the regression for non-missing data except X_1 only appears as a product with $S_{1,}$ and the switching variable S_1 also appears by itself in the regression. Note however, that the coefficients have a different interpretation and must be treated with care, as mentioned earlier.

- S_1 appears as an explicit term in the regression, although only X_1 and X_2 were originally intended to be the explanatory variables. S_1 plays an important role involving both intercepts and coefficients, and cannot be omitted from the regression without introducing major biases in the estimation of the coefficients of X_1 and $X_{2.}$ This leads to another rule that is useful when constructing graphical causal models that represent the regressions:

> Switch-inclusion rule: Every time a QDM variable such as X_1 appears in a regression, its controlling switch variable S_1 must also be included in the regression. Graphically, this is equivalent to drawing a direct causal link between S_1 and Y, even if the only causal effect of S_1 on Y was hypothesized to be via $X_{1.}$ Because of its role in switching X_1 on and off, S_1 cannot be conditionally independent of Y given X_1, but has a direct effect as well as an indirect effect through X_1. In terms of our example, *Visiting a store* has an effect on *Purchase* just because visiting enables or disables *Friendliness of the salesperson* from influencing *Purchase*, regardless of whether *Visit* has any other effect on *Purchase*, and regardless of whether *Friendliness* itself has any effect on *Purchase.*

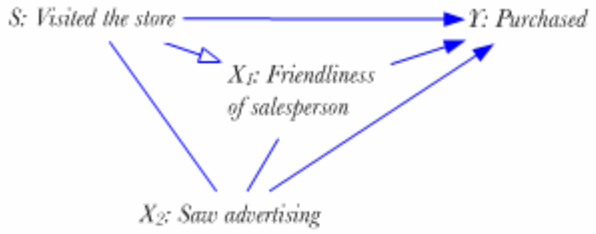


Figure 23 Example of an SEM with quasi-deterministically missing data. Arrows represent causal influence; a hollow arrow-head denotes QD switching; and undirected lines represent unanalyzed correlations.

Do we get unbiased parameter estimates with QDM data if we set up a regression comprising the terms identified in Equation 3? Although S_1 is a regular explanatory variable, we can also view it as a `missing-indicator' variable that represents the missingness of X_1, and therefore replicate the derivation used by Jones [1996] to compute the magnitude of bias in missing-indicator regression models. Using Theorem 3.1 in that paper, we know the bias to be either an additive term $\beta_1 P^m S_{12}^m F$ or a multiplicative term ($1 -$ $P^m S_{12}^m F$), depending on the parameter, where β_1 is the parameter associated with the missing variable X_1; P^m is the proportion of missing values in $X_{1,}$ i.e., P^m = 1 - mean (S_1); S_{12}^m is the sample covariance of X_1 and X_2 for the missing cases in X_1; and F is a function of the means, variances, and covariances of X_1 and X_2. In our case X_1 is missing not just in the sense that some valid value has been omitted in the dataset, but in the sense that X_1 is inapplicable and nonexistent (deterministically missing, instead of stochastically missing). Therefore the covariance term S_{12}^{m} for the missing cases is zero, and the bias terms mentioned above disappear. Thus, the regression yields unbiased parameter estimates, a fact that was confirmed experimentally via simulation tests.

Generalizing the above approach to handle QDM models with multiple switching variables S_1, S_2, etc., is straightforward, as can be proved by enlarging Equation 2 to create a separate expression for each combination of switch-values, and then deriving the analog of Equation 3. In the resulting regression, each switch variable must appear as a separate term by itself, and as interactions (product terms) of the switches. Each missing variable must appear with its respective switch, as a product term. Every variable that is *not* controlled by one of the switches must appear both with and without that switch, unless we can assume that the particular switch will not have a big effect on the behavior of the explanatory variable. The last assumption is often made because the number of terms can otherwise grow exponentially with the number of switches. Similarly, the inclusion of other interaction terms between the switched or unswitched explanatory variables is a tradeoff between complexity and accuracy, and is a general modeling consideration that is independent of the quasi-deterministic nature of the data. Care is required in the interpretation of results as mentioned earlier: the coefficient of a switched variable (i.e., of the product term) is interpreted as the effect of the switched variable in the situation in which the switch being "on"; by definition, the variable has no effect when the switch is off.

4.1.4.4 Calculating effect sizes in QDM models

Since the regression of Equation 3 only produces the "submodel" effect of X_1 as the coefficient of X_1S_1 (i.e., the effect of X_1 on Y when the switch S_1 is on) we need an additional computation to derive the "full-model" effect. The latter is of particular practical importance because it is reflects the normal manner of specifying real-world interventions. Using our previous example, a business strategist may ask: if I spend a million dollars in training salespeople to appear friendlier to customers, what are the effects on sales? The fact that friendliness only kicks in when customers *do* visit a store is half the story. The business strategist wants the total effect of friendliness, across all customers, since that's the return from the in-store intervention.

Using Equation 2, the effect of a 1-unit intervention on X_1 can be calculated as b_1' when $S_1 = 1$, and zero when $S_1 = 0$. Thus the net effect on Y of a 1-unit intervention on X_1 is b_1' times Prob($S_1 = 1$), i.e., b_1' mean(S_1), since S_1 is binary. Thus the coefficient of X_1S_1 obtained from the regression of Equation 3 must be proportionally weighted by the number of non-missing cases to obtain the total effect. This makes intuitive sense; the in-store intervention of our example would have a lower net effect if there are fewer visitors to influence. A similar derivation shows that the net effect of the non-QDM variable X_2 is $b_2'' + \left(b_2' - b_2''\right)\text{mean}(S_1)$, i.e., we use the main-effect coefficient of X_2 and add the interaction-term coefficient of X_2S_1 weighted by the switch. Thus all the information needed to compute interventional effects with QDM data can be obtained from a single regression (Equation 3) plus knowledge of the mean values of the switch variables.

4.1.4.5 Treatment of QD data for multiple regression: linear and logistic models

The previous two sections focused on quasi-deterministically *missing* (QDM) data, i.e., one of the variables was nonexistent (i.e., could not have any meaningful value, even an artificial or theoretically meaningful value) as consequence of a switch variable being "off". The more general case of quasi-deterministic (QD) data is when the dependent variable does have a value, a constant, as a result of the switch being off. For simplicity, we rewrite Equation 1 as:

Equation 4

$$Y = \begin{cases} b_0^{'} + b_2^{'} X_2 + e^{'} & \text{if } \; S = 1 \\ b_0^{''} & \text{if } \; S = 0 \end{cases}$$

where we have dropped the error term in the lower expression to indicate determinism, and retained X_2 in the upper expression as a placeholder for the usual set of explanatory variables in multiple linear regression. To avoid confusion, note that Y is quasi-deterministically dependent on S in this equation, in contrast with Equation 2 where Y had a regular stochastic dependency on S (in that equation it was X_1 which was quasi-deterministically related to S_1). Also, in Equation 4, X_2 is <u>not</u> quasi-deterministically missing when $S = 0$ although X_2 is absent in the lower half of Equation 4. The reason for the absence of X_2 despite having valid data in all cases, is because Y ignores all explanatory variables other than S when $S = 0$. To use our running example, if S represents *Visited the retailer*, and Y represents *Purchased*, it follows that Y is always 0 (did not purchase) when $S = 0$ (did not visit), regardless of any other explanatory variables X_2 (e.g., *Saw advertising*).

There are several ways to handle this situation. In the simplest case when there is only one such QD relationship in the entire model and the dependent variable Y happens to be the primary outcome of the model, the situation can essentially be ignored. That is, all cases in the dataset corresponding to $S = 0$ are deleted; then Y is regressed on all the explanatory variables with the exception of the switch S itself. The resulting submodel is substituted into the upper half of Equation 4. The constant in the lower half of Equation 4 is merely read from the dataset, and the entire equation is presented as the `nonlinear regression' explaining the primary outcome.

There are two disadvantages with the above approach, both of which relate to fact that we are dealing with a structural model comprising chains of cause and effect, not a one-stage regression. First, the above approach prevents us from simultaneously building an explanatory model for S, i.e., regression S on other explanatory variables. To use our running example, we cannot simultaneously build a model of the factors influencing visits, because we discarded all cases corresponding to shoppers who didn't visit ($S = 0$). Thus we have to solve one regression at a time, first creating a model for Y (*Purchase*), and then another one for S (*Visit*); we cannot utilize the SEM approach of maximizing overall model fit across all regressions simultaneously. Second, although it was appropriate for Y, the deletion of cases is wasteful for downstream regressions that explain other intermediate variables, since the causal mechanisms behind those variables may have nothing to do with *S*. Thus if we want to predict the effect of X_2*: Advertising* on X_3*: Intent to visit*, the deletion of cases corresponding to $S = 0$ is an unnecessary reduction of sample size, and may also introduce estimation bias because of the relationship between X_3 and *S*.

Therefore we chose to follow another simple approach that we called "reconstitution" to handle quasi-deterministic relationships in which there is a meaningful deterministic value (i.e., when the switch is off in Equation 4, the controlled variable *Y* has a constant value b_0'') that can be logically deduced from an understanding of the deterministic mechanism. In this approach we manually substitute the constant value into the dataset, if it has not been collected for that particular respondent. For example, when respondents answered "No" to F_1*: Did you receive an advertising flyer from Retailer A*, the next question F_2*: Did you look through any flyer from Retailer A* was skipped. Although the dataset now has missing values on F_2 whenever $F_1 = 0$, it is perfectly meaningful to set the value of F_2 to "No" rather than treating F_2 as a QDM variable. The reconstituted-F_2 variable no longer has missing values (other than normal sources of missingness such as respondent errors). Of course the reconstituted-F_2 variable is not linearly dependent on F_1 because of the two-part nature of their relationship: not receiving a flyer guarantees that the shopper will not look through a flyer, but receiving a flyer may or may not result in looking through the flyer.

Although QD relationships are by definition nonlinear, sometimes a linear regression ignoring the underlying determinism works reasonably well. We conducted simulation experiments in which we built a model using a single linear regression and contrasted it with a two-part model with case-deletion as described earlier in this section. The results showed that parameter estimates were often quite similar, and therefore a single linear regression using reconstituted variables is a tolerable first-approximation when it is tedious to build two-part models.

However a single regression (utilizing reconstituted variables as necessary) fails to converge when the stochastic part of the QD relationship needs a logistic regression or a similar generalized linear model (GLM), because one of the parameters in the resulting single-regression tends to infinity. For example, in discrete choice models involving categorical outcomes such *Purchased* we would write the equivalent of Equation 4 using a logistic regression component

Equation 5

$$\begin{array}{ll} \operatorname{logit}\big(P(Y)\big) = b_0' + b_1' X_1 + e' & \text{if } \; S = 1 \\ Y = 0 & \text{if } \; S = 0 \end{array}$$

Rewriting Equation 5 as

Equation 6

$$P(Y) = \operatorname{logit}^{-1}\big(b_0' + b_1' X_1 + e'\big) \cdot S$$

shows why we can't just extend our approach for linear regression and run a logistic regression on reconstituted variables:

Equation 7

$$P(Y) = \mathrm{logit}^{-1}\left(b_0^{'} + b_1^{'} X_1 + b_2 S + b_3 X_1 S + e^{'}\right)$$

Since the underlying determinism implies that *P(Y=0) = 1* whenever *S = 0*, the coefficient of *S*, namely b_2 in Equation 7 must be infinity because of the exponentiation involved in the *logit* function. (A similar effect occurs if we use the *probit* function). This is intuitively meaningful: we can think of deterministic explanatory variables (like *S*) being in some sense infinitely more powerful than the regular stochastic explanatory variables in predicting the outcome, since *Y*=0 when *S*=0 with absolutely no uncertainty. Therefore their effect size or logistic coefficient is infinitely larger than the other coefficients.

The above discussion of logistic regression with QD data also applies to more general GLM relationships, and to QDM data. For example, we can write the analog of Equation 3 as:

Equation 8

$$P(Y \mid X_1, X_2, S_1) = \begin{cases} \mathrm{link}\left(b_0^{'} + b_1^{'} X_1 + b_2^{''} X_2 + e^{'}\right) & \text{if } \; S_1 = 1 \\ \mathrm{link}\left(b_0^{''} + \qquad\qquad b_2^{'''} X_2 + e^{''}\right) & \text{if } \; S_1 = 0 \end{cases}$$

where *link()*is the appropriate GLM link function such as *logit^{-1}*. In reducing this to a single-regression approach, we expect to encounter the same nonconvergence problem because the switch variable S_1 is outside the link function.[38]

4.1.4.6 Covariance properties of QD relationships

It is important to extend the above understanding of the behavior of QD functions in one linear regression to larger structural models comprising many regressions that result from long chains of cause and effect. In particular, their effect on the covariance matrix of the observed variables has to be well understood if we wish to employ structural equation modeling (SEM) tools and techniques to estimate the model's parameters. This is important since QD relationships are inherently nonlinear, but most SEM tools only support linear models.

Our initial exploration suggests that at least for QD relationships in which the switch variables are binary and the stochastic component is linear, there exists a data transformation that permits linear SEM analysis of the transformed data, provided that the resulting are carefully interpreted. In particular, the resulting structural parameters associated with switched variables must be carefully translated back to the semantics of original variables, and the standard errors must be corrected.

To develop this understanding we begin with some basic covariance properties of QD relationships. For simplicity we assume a explanatory variable *S* causes two other

[38] It is not clear whether it is possible to design any link function that transforms a GLM such as Equation 8 into a linear regression with interactions terms.

explanatory variables X_1 and X_2 to be missing when $S = 0$; a fourth explanatory variable Z is a regular unswitched stochastic variable and used to illustrate how switched and unswitched variables relate to each other; S, Z and Y have no missing data and have ordinary stochastic relationships with each other; S is binary, but the other variables are not:

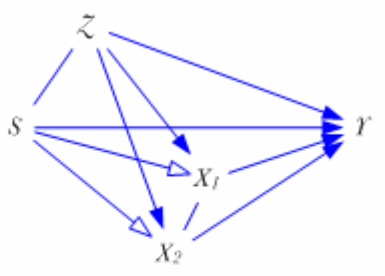


Figure 24 Example QD model for covariance analysis

The regression of Y on the other variables can therefore be written as follows:

Equation 9

$$Y = \begin{cases} b_0^{'} + b_1^{'} X_1 + b_2^{'} X_2 + b_3^{'} Z + e^{'} & \text{if } S = 1 \\ b_0^{''} + \qquad b_3^{''} Z + e^{''} & \text{if } S = 0 \end{cases}$$

Since S is a 0-1 variable, we can infer:

$$S^2 = S$$

$$\mathrm{Var}(S) = \mathrm{E}[S]\,(1 - \mathrm{E}[S])$$

Since X_1 and X_2 are missing whenever S = 0, we can infer

$$\mathrm{Cov}(X_1, S) = 0 \quad \text{if } S = 1, \text{ and undefined when } S = 0. \text{ Similarly for } X_2.$$

The data transformation that we propose utilizes the Switch-product rule (Section 4.1.4.3), i.e., we intend to replace every switched variable X_1 which contains missing values with the product variable X_1S which contains no missing values. Therefore we need to understand the properties of the new product variables. We will use the subscript "|S=1" to refer to the computation of a statistic for only the cases for which S=1. Thus Var(X_1S) represents the variance of X_1S computed across all cases in the dataset, but $\mathrm{Var}_{|S=1}(X_1S)$, is the variance computed only using the non-missing cases. Then we can derive:

$$\mathrm{E}[X_1 S] = \mathrm{E}_{|S=1}[X_1]\,\mathrm{E}[S]$$

$$\mathrm{Var}(X_1 S) = \mathrm{Var}_{|S=1}(X_1)\mathrm{E}[S] + \mathrm{E}^2_{|S=1}[X_1]\mathrm{Var}(S)$$

$$\mathrm{Cov}(S, X_1 S) = \mathrm{E}_{|S=1}[X_1]\,\mathrm{Var}(S)$$

$$\mathrm{Cov}(X_1 S, X_2 S) = \mathrm{Cov}_{|S=1}(X_1 X_2)\,\mathrm{E}[S] + \mathrm{E}_{|S=1}[X_1]\mathrm{E}_{|S=1}[X_2]\mathrm{Var}(S)$$

$$\mathrm{Cov}(Z, X_1 S) = \mathrm{Cov}_{|S=1}(Z, X_1)\,\mathrm{E}[S] + \mathrm{E}_{|S=1}[X_1]\,\mathrm{E}[S]\left(\mathrm{E}_{|S=1}[Z] - \mathrm{E}[Z]\right)$$

$$\mathrm{Cov}(Y, X_1 S) = \mathrm{Cov}_{|S=1}(Y, X_1)\,\mathrm{E}[S] + \mathrm{E}_{|S=1}[X_1]\,\mathrm{E}[S]\left(\mathrm{E}_{|S=1}[Y] - \mathrm{E}[Y]\right)$$

The above properties suggest an interesting simplification. If we zero the means of the switched variables X_1 and X_2 before we take their products with S, i.e., define

Equation 10

$$X_1^{'} = (X_1 - \mathrm{E}[X_1])S\,,$$

then we can rewrite the above properties as follows:

$$\mathrm{E}[X_1^{'}] = 0$$

$$\mathrm{Var}(X_1^{'}) = \mathrm{Var}_{|S=1}(X_1)\mathrm{E}[S]$$

$$\mathrm{Cov}(S, X_1^{'}) = 0$$

$$\mathrm{Cov}(X_1^{'}, X_2^{'}) = \mathrm{Cov}_{|S=1}(X_1 X_2)\ \mathrm{E}[S]$$

$$\mathrm{Cov}(Z, X_1^{'}) = \mathrm{Cov}_{|S=1}(Z, X_1)\ \mathrm{E}[S]$$

$$\mathrm{Cov}(Y, X_1^{'}) = \mathrm{Cov}_{|S=1}(Y, X_1)\ \mathrm{E}[S]$$

Note that the covariances of the transformed variables are identical to the covariances of the original variables (using listwise deletion whenever switched and unswitched variables are mixed) except for a scaling factor E[S]. The resulting diagrams and covariance matrices are illustrated below:

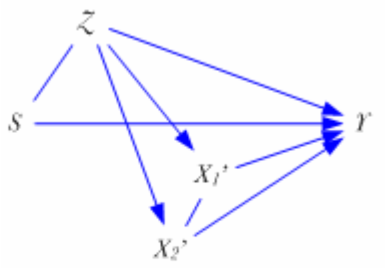


Figure 25. The example QD model with transformed *X* variables

	Original						Transformed				
	S	X_1	X_2	Z	Y		S	$X_1^{'}$	$X_2^{'}$	Z	Y
S	.	.	.	.	.	S	.	0	0	.	.
X_1	.	.	.	.	.	$X_1^{'}$	0	s	s	s	s
X_2	.	.	.	.	.	$X_2^{'}$	0	s	s	s	s
Z	.	.	.	.	.	Z	.	s	s	.	.
Y	.	.	.	.	.	Y	.	s	s	.	.

Figure 26 Comparison of covariance matrices before and after transformation
Dots represent unchanged elements; *s* represents scaling by E[*S*].

Note that the submatrix corresponding to just the switched variables X_1 and X_2 is completely unchanged since scaling all elements of a covariance matrix by a constant factor means that the implied SEM models are identical. This makes intuitive sense:

since the switched variables X_1, X_2¸etc. are nonexistent when the switch S is off, whatever we conclude about their interrelationships from studying just a submodel comprising those variables (which means studying only the cases for which data is not missing, i.e., S=1) should remain unchanged if we study the entire model using transformed variables. Put another way, studying the entire dataset using the transformed variables X_1', X_2'¸will yield the same submodel for the X variables that we would obtain by deleting all the missing cases in the dataset (S=0), and studying the untransformed X variables.

However, the relationships between the X variables and the unswitched variables Y and Z are changed after transformation; the change is a simple one, since same constant is used to scale the changed covariances. The interpretation is again intuitive: switching off the X variables part of the time means that their links to the other variables must be weakened to some degree. E.g., if S represents *Visited the store* and X_1 represents *Talked to a salesperson*, the ability of the salesperson to influence *Y: Purchase* is weakened if fewer shoppers visit the store. The degree to which the relationship is weakened is proportional to the amount of time the variable is switched off, hence the scaling by E[S].

The disappearance of the causal link between S and X_1 in Figure 25 may at first glance appear puzzling, and is related to the zeroing of the mean of X_1. Note that the original causal link in Figure 24 was of a special type: S "turned on or off" X_1, and had no further effect on $X_{1.}$ Further reflection reveals that there is no meaning to the "effect size" or structural coefficient of S on X_1 because X_1 is nonexistent when S = 0; converting a variable from a valid value to a nonexistent value cannot be measured in any "units of change". Therefore there was no parameter to be estimated on the structural link between S and X_1 in Figure 24; the link only represented the genuinely causal effect that S has in making X_1 nonexistent. Similarly, the relationship between X_1 and X_2 in Figure 24 had a special meaning: it reflected the structural link between the two variables *when both exist*, i.e., when S = 1. These peculiarities originating from the quasi-deterministic switching of the original X variables in Figure 24 have correctly disappeared when we changed to the transformed variables X_1' in Figure 25. In the transformed model, all the variables exist all the time; there is no switching; and a standard structural equation model can be estimated over all the variables.

It is worth revisiting Equation 9 and examining the corresponding regression in the transformed model to find out whether parameter estimates have been biased. The analogous regression based on Figure 25 is

Equation 11

$$Y = a_0 + a_1 X_1' + a_2 X_2' + a_3 Z + a_4 S + e$$

We compute the regression coefficient a_1 as follows [Spirtes et al. 1998]:

$$a_1 = \frac{\mathrm{Cov}(Y, X_1' \mid X_2', Z, S)}{\mathrm{Var}(X_1' \mid X_2', Z, S)}$$

Now $X_1' = 0$ whenever S = 0 by virtue of its definition in Equation 10, and hence $\mathrm{Cov}(Y, X_1' \mid ...) = 0$ whenever S = 0. Therefore we can write $\mathrm{Cov}(Y, X_1' \mid ..., S) = S \cdot \mathrm{Cov}_{|S=1}(Y, X_1' \mid ...)$ where we have moved the conditioning by S = 1 to the subscript using the notation introduced earlier. It follows that $\mathrm{Cov}(Y, X_1' \mid X_2', Z, S) = S \cdot \mathrm{Cov}_{|S=1}(Y, X_1' \mid X_2', Z) = S \cdot \mathrm{Cov}_{|S=1}(Y, X_1 \mid X_2, Z)$ since $X_1' = X_1$ when S = 1 (the means drop out). A similar derivation shows that $\mathrm{Var}(X_1' \mid X_2', Z, S) = S \cdot \mathrm{Var}_{|S=1}(X_1 \mid X_2, Z)$. Therefore we conclude that

$$a_1 = \frac{S \cdot \mathrm{Cov}_{|S=1}(Y, X_1 \mid X_2, Z)}{S \cdot \mathrm{Var}_{|S=1}(X_1 \mid X_2, Z)} = b_1' \text{ from Equation 9.}$$

In other words, the regression coefficient of the transformed variable *X₁'* in Figure 25 is identical to the coefficient of the original *X₁* variable in Figure 24 (assuming that we only used the data cases for which the variable was not switched off). This is similar to our inference in Section 4.1.4.3 that we can use the transformed variable in place of the original QDM variable provided that we are careful to interpret the resulting coefficient as the effect of *X₁* in the *submodel corresponding to the switch being on.* Thus the coefficient does *not* represent the effect of *X₁* in the full model in which the switch *S* may be either on or off; this coefficient must be weighted by E[*S*] to correctly obtain the full-model effect size.

A similar derivation shows that the coefficient of *Z* in Equation 11 works out to its equivalent in Equation 9.

$$a_3 = \frac{\mathrm{Cov}(Y, Z_1 \mid X_1', X_2', S)}{\mathrm{Var}(Z, \mid X_1', X_2', S)} = b_3' S + b_3''(1 - S)$$

In other words, the replacement of the original *X* variables by their transformed counterparts *X'* did not alter the effect of *Y* on *Z* (net across both submodels corresponding to the switch being on and off). This implies that the coefficients of unswitched variables in the transformed dataset can be read directly from the regression results without correction.

A final derivation shows that $a_4 = b_0' - b_0''$ indicating that the coefficient of *S* can be directly read from the regression using the transformed dataset as the net direct effect of *S* on *Y*. Although by definition a "direct effect" excludes effects along other paths from *S* to *Y*, the coefficient does include the effect that *S* causes in *Y* by turning the *X* variables on and off (as illustrated by the difference $b_0' - b_0''$). Note that the two statements are not contradictory; there is no path from *S* to *Y* that passes through the *X* variables in Figure 25; the manipulation of *Y* by quasi-deterministically switching *X* does not get rendered as a separate path in the path diagram, and the effect size is absorbed into the direct link between *S* and *Y*, instead of being routed through *X*.

The above derivations show that the transformed model of Figure 25 produces intuitively correct parameter estimates. However one caveat must be observed: the standard errors and significance tests can be affected for parameters that involve the transformed variables. The reason is that we only obtained data on *X₁* when *S* = 1; however, the transformed variable *X₁'* has data for all cases, because the multiplication by *S* resulted in *X₁'* getting populated by zeros whenever *X₁* was missing. This artificially inflates the sample size for those variables in a way that standard SEM tools are unaware of, since current tools are unable to handle missing values and must be fed the transformed dataset.

The corrections required are complex since they depend on the combination of variables chosen: e.g., for a coefficient linking X_1 and X_2, since both variables are switched by S, the sample size is effectively the number of cases when the switch was on, i.e., N E[S]. However, for a coefficient linking X_1 and Z, the effective sample size may be larger, and tools are unavailable to calculate precise significance levels. We relied on qualitative judgments, i.e., being extra conservative in terms of acceptable p level cutoffs and narrow confidence intervals to draw inferences from our data.

Another point to remember when interpreting and reporting results is that any inference involving the actual value of the transformed variables requires a mapping back to the original scales, because the transformed variables had their means zeroed.

In summary, the "mean-normalized switched-product" (MNSP) transformation technique that we presented in this section transforms a dataset governed by quasi-deterministic relationships into a standard dataset governed by ordinary linear relationships. With some extra care in the analysis and interpretation of results, QD models can be handled quite effectively using common SEM tools.

4.1.4.7 Modeling indifference and inattention: Special classes of QD relationships

The examples of quasi-deterministic relationships that we have provided so far (especially in Section 3.2.3) were fundamental structural relationships that we discovered in the particular subject domain of shopping for a product; there was no particular theme that connected one example of a QD relationship with another example (other than the very fact that they were QD relationships). However, there are two particular classes of QD relationships that appear repeatedly in surveys, and are of sufficient generality and importance to merit being called out. These are the "Don't Care" and the "Didn't Think about It" classes.

Good psychometric design requires that respondents are not forced to provide a response to questions that they really can't answer. In particular, survey questions must avoid making implicit assumptions that may not be true for all respondents. For example, if the study theorizes that shoppers may visit a retailer because the retailer is known for high-quality products, it is insufficient to simply measure the shopper's perception of the retailer's product quality. Although the shopper may very well *have* an opinion about the retailer's quality at the time of answering the survey, the shopper may not have *had* the perception at they time they visited the retailer. Or even if the shopper had the perception of quality, they may not have thought about it before visiting the retailer, and therefore the perception may not have played a role in inducing the visit.

A survey that just asks for the respondent's perception on a scale of 1 to 5 elicits the respondent's current views and loses crucial causal information that could indicate that the perception is irrelevant for influencing shopper behavior. Worse, the respondent may have never had an opinion about this particular topic throughout the entire purchase process, both before and after visiting the retailer. However being confronted with a survey question precipitates the formation of an opinion after the fact; when forced to respond with an answer on a 1 to 5 scale, people pick something and move on. The answer may be random or may be systematically influenced by something that happened recently, e.g., post-purchase experiences with the product. This introduces noise into the analysis and possibly systematic biases, depending on how the perception was constructed. Instead, it is desirable to directly capture the fact of "not thinking about" the topic, infer a lack of causal influence from the fact, and avoid the collection of an irrelevant perception from the respondent. This is done by adding a mutually exclusive option similar to the "Don't remember" option:

Please think back to what you thought about the following statements before you first visited Best Buy for TVs. If you did not think about a statement at that time, please select the `Didn't think about it' option.

	Strongly Agree	*Somewhat Agree*	*Neither Agree nor Disagree*	*Somewhat disagree*	*Strongly disagree*	*Didn't think about it*	*Don't remember/ Don't know*
I expected that the retailer would have high quality TVs							

Therefore the survey documents either the perception or the fact that the respondent didn't think about it. In subsequent analysis, the "Didn't think" response is handled quite differently from the "Don't remember" response: the former is involved in a quasi-deterministic relationship, whereas the latter is a traditional form of missing value and handled by standard techniques such as listwise deletion (see Section 4.1.8 for the latter). We generate two variables: a binary scaled *Thought-about-quality* variable with value False when respondents selected the "Didn't think" option and 1 otherwise; and the Likert-scaled *Quality-perception* itself which has systematically missing values when *Thought-about-quality* is False. In other words *Thought-about-quality* quasi-deterministically controls *Quality-perception* and has all the associated causal semantics.

Note that *Thought-about-quality* is not a "dummy" variable (missing indicator) introduced to represent the "missingness" of *Quality-perception*. On the contrary, it is a full-fledged explanatory variable with its own meaning (e.g., consider that we could build a model using just *Thought-about-quality*, discarding the perception itself) and its own causal relationships to other variables in the model (e.g., what made the shopper think about quality? How does thinking about quality affect *Visited-retailer* and other outcomes? If we make shoppers 10% more likely to think about quality---regardless of what their perception of quality itself might be---what is its effect size on sales?) The actual *Quality-perception* itself kicks in when the switch *Thought-about-quality* is "turned on". Calculating the effect of *Quality-perception* on *Visited-retailer* or sales involves the usual switched-product computation methods for QDM data described in Section 4.1.4.4.

A similar pattern of QD data arises when respondents *do* think about a particular topic, but don't really care about it. For example:

… At the time that you visited the store, how did you feel about.

	Excellent	*Very good*	*Good*	*Average*	*Poor*	*Didn't care*	*Don't Remember/*

							Don't know
The looks of that TV?							

Unlike "Didn't think about it", the notion of "Don't care" has received some attention in the decision making literature because it is a direct expression of preference data. For example, utility theory models both the value of an attribute and the weight of an attribute; the weights measure the "importance" of the attributes relative to each other. Therefore not caring about an attribute can be construed as a low weight on that attribute. While this is not an incorrect approach in principle, in practice it has some deficiencies. First, it forces us to assume a utility-based model, which we have eschewed for this study both because the assumptions of utility theory have been shown to be repeatedly violated in real-world settings, and because we favored a theory-less discovery-oriented approach. Secondly, if respondents really didn't care about the attribute, forcing them to provide a rating on the attribute just precipitates noisy data with suspect causal influence in exactly the same manner that we described above for the "Didn't think about it" class of relationships. So a utility theory based approach would employ a meaningful attribute-weight multiplied by a noisy or irrelevant attribute-value.[39]

"Don't care" or indifference has also received attention in the psychometric literature because it can be the midpoint of a bipolar scale that ranges from "Very good" to "Very bad". Mapping "don't care" to the middle of the scale may appear meaningful when the scale is considered by itself, but is problematic when considering the causal effect of the given variable on another variable. In effect, we would be asserting that if the TV had moderately good looks, the inducement to purchase would be same as if the purchaser did not care about the looks of the TV, which is far from evident. A purchaser who cares intensely about the looks of a TV and rates it moderate relative to other TVs she has seen is unlikely to purchase that particular TV, whereas a purchaser who ignores the looks of the TV might very well purchase it because the other attributes might be excellent. Mapping "Don't care" to the scale midpoint or to a 0 coding does not disable the "looks" attribute from playing a significant role for respondents who don't care about looks. In general any mapping of a "Don't care" response to the actual attribute scale itself will fail to correctly capture causal influence.

"Don't care" responses cannot be handled via missing value methods either. For example, listwise deletion results in biasing the data towards people who only care about the attribute, which would grossly inflate the importance of the attribute.

Clearly, the semantics of not caring about something must be handled separately from the actual perception of that thing, just as we described above for "Didn't think" behaviors. Thus, the data elicited for the above example was transformed during analysis into two variables: a binary variable *Cared-about-looks* and a Likert-scaled *Feeling-about-looks* variable; the former is a quasi-deterministic switch variable that causes the latter to go

[39] This problem would disappear in the special case that the attribute weight is zero, in which case there was no point in eliciting the noisy attribute rating in the first place: this would correspond exactly to our handling of the data using the concept of QD relationships.

systematically missing. Again, the *Cared-about-* variables are not dummy placeholders for missingness, but have full semantics and causal effects that are distinct from the *Feeling-about-* variables.

Incidentally, we found that there were a few attributes in our dataset (such as the looks of the TV), for which almost every respondent cared about the attribute, making the Cared-about variable nearly constant with value *True*. In this extreme case, we could just drop the *Cared-about-looks* variable from the analysis, and delete any cases in which the variable was *False*. The small error was worth the gain in simplicity: *Feeling-about-looks* could now be handled as a regular stochastic variable instead of being treated as a switched QD variable.

4.1.4.8 The importance of QD models

Experiments on our consumer behavior dataset have demonstrated the failure of logistic choice models to converge numerically when the true underlying mechanisms are quasi-deterministic. This has startling implications: we are in effect stating that as commonplace an activity as *Decision to Purchase* cannot be explained as a function of *Visit to the store* using standard discrete choice models available in the literature, despite the fact that the logic of the relationship between these variables is so intuitively evident to anyone who thinks about it for a moment. The underlying quasi-deterministic mechanisms cannot be ignored.

Some "numerical" approaches have been proposed in the literature for dealing with nonconvergence issues, such as manual introduction of noise into the data as with Ridge regression. The idea is to make the deterministic component slightly stochastic and thereby increase coefficient stability. Our view is that these approaches are inappropriate. The nonconvergence is not a spurious or chance effect as in the phenomenon of multicollinearity for which these approaches were originally designed. Rather, the underlying determinism is a true fact of nature, a mechanism that we should be glad to have discovered, and which should be established as an explicit part of the model, not to be disguised or avoided. In fact, the partial determinism can be exploited to improve model accuracy and R^2 as described below. This philosophy runs contrary to the traditional statistician's advice "the best solution to the missing data problem is not to have any!" [Allison 2001].

Quasi-deterministic models can in general be handled by estimating the two parts separately for each QD functional relationship, which is computationally tedious, especially when there are a number of QD functions within the model. As mentioned above, linear regression using switch-product terms produced a good enough approximation that we found it useful for practical purposes in our study. However it is worth noting that properly estimated QD models can provide substantially more predictive accuracy (as indicated by measures of model fit such as R^2) than such linear approximations. The reason is easy to understand: the deterministic component of a QD relationship is an inherently error-free component; it is capable of perfect prediction when $S_1 = 0$. Thus, the noise in the model can be reduced for at least part of the dataset, corresponding to those cases in which the switch is off, by explicitly separating out the deterministic part from the stochastic part before estimating of the latter, and then reassembling the two parts after parameter estimation. The prediction accuracy thus goes up, and the extent of improvement depends on how large the deterministic component is relative to the stochastic one. In the extreme case of a nearly fully-deterministic model, the switch predicts the outcome almost all the time, and R^2 is close to 1. For more realistic scenarios with a moderate mix of determinism and stochasticity as in our dataset, the improvement in R^2 is more likely to be on the order of 10-20% over a linear model that ignores the determinism.

4.1.5 Modeling the outcome variable: choice of retailer

Since there are numerous electronics retailers from whom shoppers could have purchased their TVs, the outcome was in essence a nominal variable with an unknown number of possible values. Similarly, since shoppers could and did shop at varying numbers of retailers, many of the explanatory variables had to be elicited for each retailer, e.g., *Helpfulness of salespeople at Retailer A*, *Helpfulness of salespeople at Retailer B,* and so on, since the helpfulness (or lack of it) at retailer A could affect the decision to purchase at retailer B. For pragmatic reasons, including respondent fatigue and sparse distribution of data for the smallest retailers, this repetitive elicitation of data for each retailer from each respondent had to be restricted to finite set of retailers. Pilot experiments with a full-length survey showed that asking respondents to provide data on up to 5 retailers was manageable. This posed the question of what is the best way to partition the set of all retailers in the outcome variable in a way that maximizes the usefulness of the model, especially to our client, while minimizing the biases induced by grouping a subset of retailers into a single value of the outcome variable. The bias arises not from the partitioning of the outcome itself, but from structuring data acquisition for the explanatory variables on those retailer groups, thereby inducing a backward dependency of the explanatory variables on the outcome. Thus one of the criteria for minimizing bias was to minimize the introduction of backward dependencies. Other important criteria were to maximize the amount of useful data obtained while minimizing respondent fatigue.

Clearly, we needed to query respondents about their experiences with our client, Retailer A, since we had to at least know whether or not they had made their purchase from this retailer. Similarly, we could query respondents about the next three biggest retailers B, C & D. However if respondents had not purchased from any of these “Top 4” retailers A through D, the fifth retailer about whom we acquired data necessarily had to be the one from whom they purchased, in order to avoid a missing value on the outcome variable, which in turn would have meant that the entire case would have to be discarded. This implied grouping all other retailers into a single ‘pseudo-retailer’, O. The actual fifth retailer about whom data was elicited thus differed for each respondent, and the following algorithm was use to select this retailer.

Early on during the survey, each respondent was asked for a list of all retailers that they were aware of as carrying TVs, and also the retailer from whom they had purchased. The respondent was subsequently queried about the Top 4 retailers whether or not the respondent had purchased their TV from any of these retailers. If the retailer from whom the respondent had purchased, was not one of the Top 4, the `Other Retailer O’ was automatically chosen to the retailer from whom they had purchased. Otherwise O was randomly selected from list of all retailers that they were aware of, minus A through D[40]. This ensured that all explanatory variables relating to retailers A through D, which included our client, would be bias-free. Variables relating to retailer O were positively biased, since they had a disproportionate number of purchases. These variables were therefore only useful for some qualitative inferences in which they were treated as reflecting a `pseudo-retailer' representing `the best of all other retailers'. In particular, these variables were not used in our causal model, to keep it bias free. Since retailer from whom the respondent purchased was always selected, the outcome variable had no missing data and all cases were maximally used.

Thus each respondent provided complete data on 1 to 5 retailers depending on how many they had visited. Note that after the above retailer selection algorithm decided which retailers the respondent should be queried about, the survey instrument randomized the order of retailers in which the respondent was queried, to avoid bias effects from the sequence.

This formulation resulted in an outcome variable with five nominal values, which permitted modeling its functional dependency on its parent causes via multinomial probit and logit functions. Since the primary focus of analysis was purchasing from a single retailer, namely our client, the standard transformation to a binary variable (i.e., Purchase vs. Non-purchase from Retailer A) was used. This philosophy of favoring a Retailer-A-specific model over a single multi-retailer model resulted in treating the variables that were repeated across retailers as columns in our analysis dataset, although the survey instrument treated them as rows.

Over the course of the analysis, the functional relationship between the outcome and its direct causes was studied using several formulations other than probit/logit, including a simple linear function and some nonlinear regressions based upon a causal understanding of the mechanisms involved. The latter refers to the multiplicative effect of product screening attributes (Section 4.4.3.1) and to the quasi-deterministic effect of key variables such as whether or not the respondent visited the retailer, and whether or not they seriously considered at least one TV at the retailer. Of these functional forms, the nonlinear regressions provided the best fit (as measured by pseudo R-square), the GLM models provided good fit, and the linear model provided a reasonable approximation to the other two models. For computational simplicity (since the causal discovery tools large linear at present), linear models were used in most of the semi-automated parts of the analysis methodology, and the nonlinear aspects were superimposed either by model

[40] Note that since the sampling frame comprises retailers that the respondent was *aware of*, it creates a backward dependency between the awareness-variables and the outcome. This does induce a bias in parts of the model that are causally antecedent to awareness; causally descendant portions are not affected (the implied descendant model is a simple quasi-deterministic pattern: if respondents were not aware of a retailer, they did not buy from the retailer; else they bought according to the other variables). Since only a tiny section of the model is thus affected (mainly the Prior History section), this choice of sampling frame was judged to be the most efficient.

splitting (Section 4.3.3) or by manual improvements to the auto-generated structural models.[41]

In passing it is worth noting that our study was designed to support analysis and modeling of several other useful outcome variables beyond the choice of retailer: choice of a particular (manufacturer) brand of TV, type of TV etc. Although the data that we acquired enables such analyses, for practical reasons this was subsequently excluded from the scope of the current study and are not reported further.

4.1.6 Key intermediate outcomes and model-splitting

Many of the variables that ended up in our model as direct causal drivers of the *Purchase* decision were quite ordinary and predictable variables such as *Liking for the TV, Price,* and so on. However there were several non-standard variables such as *Considered at least one TV at the retailer,* which has quasi-deterministic effect on *Purchase* (Section 3.2.3). Each of these variables in turn had several other variables as their own causes. The entire collection of hundreds of variables formed a complex graph of cause-and-effect relationships. Many of these causal chains were quite long, e.g., there were 5 intermediate variables on the shortest path between *Received a flyer* and *Purchase*¸ and many more along the other paths. There were loops in the graph due to the multiplicity of mechanisms that were simultaneously at work. Thus the entire model formed a complex picture of causal interrelationships. This picture would be simplified to some extent by the end of the quantitative analysis by dropping numerous links whose influences were quite weak, but the model remained fairly complex nonetheless, unless oversimplified to just present the `top ten' critical factors.

We emphasize the complex interrelationships within the model to contrast it with the naïve models often built in the marketing literature that attempt to impose a tree-shaped structure on the world (e.g., see our discussion of regression trees in Section 4.1.1.4). This kind of decomposition is misguided; a tree-shaped structure implies numerous conditional independences that are simply false empirically, as our model illustrates, and insisting on such false simplicity results in incorrect estimation of the effect sizes of the variables, leading to erroneous conclusions about their importance. The crisscrossing cause and effect interrelationships that we empirically observed just reflect the complexity of shopping in general.

However, we did find one distinctive point of decomposition that appears to cleanly break up the complex structure into two parts without introducing too much error. That role is played by the intermediate variable *Visited the retailer*. Note that this variable also has a quasi-deterministic influence on the *Purchase* outcome: if the shopper did not visit the retailer they definitely did not purchase a TV, but if they did, all the post-visit

[41] Our qualitative research suggested yet another functional form for the outcome, that has the potential to substantially improve model fit and predictive power, but which was not fully tested during the quantitative phase due to reasons of scope: The final choice of retailer is determined by the highest Liking-for-TV-at-a-Retailer, the best Price cross all retailers, and other comparisons., computed over the set of retailers that the shopper visited where they seriously considered at least one TV. This is a max() function complicated by the presence of quasi-deterministic skips and nonlinear interactions, and thus does not quite correspond to the propensity-to-buy latent variable formulation that underlies a traditional probit/logit formulation.

variables relating to store experience, price, product attributes, etc., kick in with some stochastic influence. Indeed if the shopper did not visit the retailer, almost half the variables in the model become quasi-deterministically missing (see Section 4.1.4 for definitions). Only the pre-visit variables remain with meaningful data. Thus we can divide the model into pre-visit and post-visit sections. The *Visited retailer* variable visually appears to be a hub at the center of the model (see the yellow node at the center of Figure 1). All the pre-visit variables directly or indirectly influence this intermediate variable, and thus can be viewed as a standalone causal model that explains the *Visited* outcome. However note that pre-visit variables can also directly influence the final *Purchased* outcome, in addition to their indirect influence through *Visited.* Thus we cannot build a model that explains *Purchased* using just the post-visit variables---unless we have license to do so via justifiable conditional independence relationships. It turns out that this is indeed the case; see our analysis in Section 4.3.3 for more on this decomposition.

Simple though the decomposition is in hindsight---after all it is quite intuitive to think about one model that explains *Visits* and another model that explains *Purchases*---it appears that this kind of structure has not been described in the literature. Part of the problem appears to be the quasi-deterministic nature of the intermediate variable, *Visited retailer*, which, as described in Section 3.2.3, statisticians strive to avoid. The other problem is that the conditional independences that allow pre-visit variables to be partially excluded from a post-visit model have to be empirically justified, as we did.

4.1.7 Power analysis

As is the case with most structural equation models, detailed power analysis for each parameter to determine the desirable sample size was impractical due to the size of the model and the presence of quasi-deterministic patterns and other nonlinearities. We applied the most commonly used heuristic: the number of cases should be at least number of latent variables times 3 items per latent variable times 5 cases per item. An initial estimate of 50 latent variables led to a sample size requirement of 1500 complete cases. Subsequent experience with the causal structure discovery methodology showed that we had significantly overestimated the number of latent variables (Section 4.1.1.2), and underestimated the reduction in data due to quasi-deterministic skips. In retrospect, the sample size turned out to be adequate for almost all parts of the analysis, but it is evident that unlike traditional confirmatory factor analysis, new power analysis methods or heuristics are needed for use with the newer exploratory modeling methods we describe. In particular, when quasi-deterministic patterns are present, a pilot sample to determine the extent of the omitted data is advisable.

4.1.8 Treatment of missing values

The handling of missing values drew considerable attention in our study because of the apparent sparseness of the data (see Section 4.1.4.2), and as we described earlier, the underlying issue turned out to be not really the “missingness” of specific variables but the quasi-deterministic relationships connecting the variables (see Sections 3.2.3 and 4.1.4). Quasi-deterministically missing (QDM) data accounted for the vast majority of missing data, and was handled either by the Switched-product method (Section 4.1.4.3) or by the

Reconstitution method (Section 4.1.4.5) depending on whether the missing cells of the variable could be logically reconstituted or whether the variable was logically inapplicable given the state of its controlling switch-variable.

While QD relationships were responsible for the vast majority of empty data cells in our dataset, there were also some missing values that arose from traditional mechanisms, and this section focuses only on such traditional missing values. Common reasons for missing values were survey respondents forgetting to answer a questions, not knowing the answer (say because their spouse did that part of the shopping and wasn't available supply the answer), the question was sensitive (e.g., race or income), and so on. The number of such cases was generally quite small (under 5%) per variable, except for the rare difficult question that taxed the respondent's memory (about 30% missing due to a "Don't remember" response). The variables with the most missing values also often had low explanatory value, and were sometimes dropped for other modeling reasons.

Allison [2006] provides an excellent review of the analytical procedures appropriate for handling missing values. In general, complete-case analysis (aka listwise deletion) is a superior approach from the viewpoint of avoiding estimation bias, its main problem being the reduction in sample size since missingness accumulates over all the regression variables we choose to introduce. Pairwise deletion is unsatisfactory because the resulting inconsistent sample sizes of the variables leads to incorrect computation of standard errors and other test statistics. Fixed value imputation methods, e.g., substituting a variable's mean, median or other fixed value (computed using the non-missing data cells) into the variable's missing data cells generally produces estimation bias. The missing-indicator methods (aka dummy variable methods) either produce bias, or when bias-free provide no advantage over complete-case analysis while adding complexity. Multiple imputation and maximum likelihood estimation are better approaches; they are not biased under MCAR but they do at least require the MAR assumption. These two approaches are also a lot more complex and need specialized software; we did not deem them worthwhile for dealing with the small number of traditional missing values in our dataset. Since complete-case analysis is unbiased under the MCAR assumption for the dependent variable (missingness of Y is not related either to Y or to e) and robust even to violations of the MAR assumption for the independent variables (i.e., missingness of X is unrelated to Y), it was considered the best technique for most of our regression analyses. Additional care was taken to increase the likelihood of satisfying these assumptions. In particular, we explored theories on what caused the missingness of the dependent variable Y. If missingness appeared to be due to one of the covariates, that was deemed acceptable, but if missingness was due to another variable that also correlates with Y, we added the other variable as a new covariate in the regression to avoid bias.

4.1.9 Modeling sequence effects

During our qualitative research, we observed some distinct sequence patterns in shopper behavior, e.g., shoppers were more likely to visit Best Buy before visiting other electronics retailers, and shoppers were likely to compare TVs seen at later stores to specific TVs seen at earlier stores. It is obvious that if Circuit City is typically visited after the shopper has gone to two other retailers, and Sears is typically visited after three other retailers, then Sears has overall a lower chance of making a sale than if the order were reversed, since the shopper might find a satisfactory deal at the earlier retailers, or get tired of shopping after visiting the earlier retailers, and so on. Therefore we hypothesized that the sequence in which shoppers visited stores might be important for explaining outcomes such as market share.

The problem becomes more complex if we recognize that shoppers make multiple trips to each retailer, often visiting other retailers between returning to an earlier retailer. There are history-dependent path effects: a good salesperson at an earlier store may clinch a deal whereas a lousy shopping experience may induce the customer to visit another store. The whole issue of temporally structuring “shopping trips”, “store visits”, etc., is a complex one, and some kind of a temporal model (e.g., a state-transition model) appears to be desirable. The challenge is to integrate such temporal modeling techniques with standard causal models of the kind we’ve described earlier. Recall that a fundamental requirement of any model we produce is to be able to predict the causal effect of an intervention, not to just make a prediction given an observation (Section 4.1.1), and traditional techniques such as hidden Markov models are not necessarily causally consistent.

Given that we obtained excellent explanatory power (R^2 of around 0.7) with a standard causal model, and given a guesstimate based on inspecting the data that sequence effects might improve that by 10% we decided that for the purposes of the current study that the complexity of sequence modeling was not worthwhile; however it does remain an intellectually interesting question. Note that we did account indirectly for some temporal effects because we had feedback loops in the model; e.g., *Visited Circuit City* had a negative effect on *Visited Sears*. However, we did not explicitly incorporate sequence data obtained via direct questions in the survey.

4.2 Study design and data collection

While the Discovery Phase empirical research was the primary source of input (i.e., the variables and the causal mechanisms) used for developing the quantitative model, we also utilized two secondary sources: the prevalent theories in use at our Client, and the academic literature. It is important to include the former even when the qualitative research suggests that the client’s beliefs are incorrect, because there are usually people in the client’s organization who resist and disbelieve the results unless presented with the most solid evidence. This requires explicit incorporation of the client’s hypotheses into the model and testing at statistically significant levels. Unless presented with such evidence, it is natural for a client to persist in believing that their own folk theories are “also correct”. Also, the results of a study often seem “obvious” in hindsight, because people forget that the hypotheses finally found to be true were among numerous other hypotheses also thought to be true at the beginning of the study and the client didn’t actually know which variables really mattered. Therefore it is important to establish a baseline of prior beliefs that can be contrasted with the findings from by the study, and we conducted a series of structured exercises with experience executives from our Client to elicit their hypotheses. For example they believed that the Top 4 critical factors that drive sales were (1) Robust assortment (2) Carrying the right brands (3) Having the item in stock (4) Value. We ensured that our model contained variables capturing these factors, and it is worth noting that only one of these four eventually turned out to be critical, and in a different form than hypothesized: Perception of Getting a Good Deal.

The marketing literature provided scant assistance in augmenting our qualitative model since thorough empirically-grounded models of causal drivers of sales are hard to find, and most papers reviewed generated their constructs either theoretically or via weak empirical research. Most useful was the handbook of marketing scales [Bearden and Netemeyer 1999] which we used to design specific items whenever the corresponding constructs were present in our model.

In order to obtain the data required to build the quantitative model, we chose to use a retrospective survey of people who had recently purchased a television set. This decision was based on several factors, including the difficulty in obtaining real-time data on variables spanning a protracted shopping process and the definition of the outcome variable. Since our primary outcome was the consumer's choice of Retailer A versus other retailers, rather than choice to buy a TV versus not-buy a TV, we did not need to recruit people who decided not to buy a TV.

Our sampling process is described below in Section 4.2.1, and the construction of the survey in Section 4.2.2. The final survey contained about 150 main questions, yielding about 1500 variables because many questions had multiple parts, and because the questions were repeated for up to five retailers as described in Section 4.1.5. The survey covered a very broad range of areas that could potentially influence the purchase outcome, including the shopper's history of prior experiences with that retailer (e.g., awareness of the retailer, shopping at the retailer, prior purchases), advertising seen, store location and convenience, store characteristics, salespeople, product attributes, pricing, promotions, and demographic information about the shopper.

While the sampling process was telephone-based, we had many choices on how the survey could be administered to respondents, including an online survey, a telephone-based survey, or a paper mail-based survey [Nathan 2001, Miller and Dickson 2001, Krosnick and Chang 2001, Berrens et al. 2003, Ilieva et al. 2001]. We chose online surveys as our primary method, because this had significant advantages in terms of uniformity of administration (avoiding administrator bias), and lower costs. The main disadvantage, viz., not all respondents have easy access to the Internet, was ameliorated by using telephone administration for those who refused or could not participate online. In the end, about 20% of respondents provided their answers over the telephone, and the rest filled in the online survey. When we checked the survey answers we did not find any pattern of differences between online versus telephone respondents.

The survey was programmed using a commercial online survey administration tool, Confirmit. Programming the survey posed a significant challenge due to the extensive presence of quasi-deterministic relationships between the variables, which translated into a number of skip patterns and text substitutions based on the structure of the relationships. For example, if a respondent did not visit a particular retailer's stores for TVs, but did visit the retailer online, the questions pertaining to their experience with the retailer's salespeople were skipped; but the questions relating to price and product attributes were retained. Similarly, since many questions depended on a separation between the time periods before and after visiting a retailer for reasons of establishing causal sequence (e.g., if the respondent read a flyer after visiting the retailer, it would not be treated as a potential cause for visiting the retailer), the text of each question had to be automatically modified to reference the period before visiting, whenever the respondent had visited the retailer. The large number of logical skips, substitutions, and sequencing constraints, which substantially improved the fidelity of the data by the resulting careful step by step reconstruction of the respondent's experience, thus incurs an unavoidable additional cost in survey programming and testing; indeed an activity typically expected to take a week's effort took over a couple of months to program.

Respondents were provided a unique ID and unique Web link, which served as an authentication mechanism to limit access only to recruited respondents, and also enabled us to identify and track the respondent when they came online to fill in the survey.

Setting up the survey took about 2 months because of the need to program and test the quasi-deterministic logic, and administration of the survey took about 3 months because of the low incidence rate. About 30 survey responses were discarded due to inconsistencies or recruiting errors, yielding our final sample size of 1504. The data was then analyzed as described below in Section 4.3 to build the quantitative causal model, estimate the effect of each variable on sales, and identify the top ten critical factors.

4.2.1 The sampling frame and screener design

A national (U.S.) sample of 1504 recent TV purchasers was recruited for the study. Telephone-based random digit dialing (RDD) was used to locate this sample, since this is considered to be the `gold standard' of sampling methods, used almost without exception in medical and legal studies. Although it is standard practice in market research to use less representative samples (e.g., to focus on a few a major cities, or to use pre-recruited panels) because of their lower costs, we did not want to compromise the trustworthiness of our model due to sampling practices considered questionable by some researchers.[42] Therefore we obtained EPSEM RDD samples from reputable vendors such as Genesys, STS, and SSI.

Since our client's interest was in studying the purchases of high-end TVs, we excluded all purchases of TVs under 15 inches in size, all "regular TVs" under 27 inches, and all "front projection" TVs which are likely to induce a completely different purchasing process.

"Recent TV purchasers" was defined as households that had purchased a TV during the previous 8 months. This period was a compromise between the increasing likelihood of recall error with people who had purchased too long ago, and the serious difficulty in finding households who had purchase TVs. Based on the screener, we estimate that about 10% of all households qualified in terms of meeting the recruiting criteria. About 60% of the households that qualified agreed to participate in the survey. Thus the overall incidence rate was about 6%, which from a survey research perspective is a tiny fraction of the population, difficult to reach via random digit dialing and therefore expensive to recruit. We estimate that we dialed over a million phone numbers (of which about 60% percent turned out to be disconnected numbers) in order to recruit our 1500 respondents.

[42] But see our discussion of alternative approaches in Section 4.2.4

4.2.2 Survey design and survey length

We had several criteria in mind when designing the survey. First and foremost, we desired high fidelity with respect to the variables in the qualitative causal model, i.e., that the data obtained accurately reflected the constructs identified in the earlier Discovery Phase. Our idealized conception of the survey was that of a `scaled-up virtual ethnographer', i.e., we imagined a device capable of capturing subjects' behavior at the extremely detail-rich and fine-grained level that we executed manually during the Discovery phase, yet capable of obtaining this data on a sufficiently large scale to support rigorous statistical analysis. However, for the practical reasons described earlier we were limited to a retrospective survey, and therefore we faced several design problems: how do we minimize problems with recall when some of the respondents had completed their shopping several months earlier? How do we deal with language issues when some respondents do not speak English well? Given the large size of our model, how do we obtain the maximum amount of data possible from each respondent with the minimum amount of subject fatigue? Another criterion was to minimize the total cost of the operation, including costs from recruiting, dropouts, incentives, and data processing. Clearly these criteria are related to each other, e.g., a larger survey size can result in lower participation rates and more dropouts [Miller and Coates 2003].

There was considerable concern about the length of the survey and respondent fatigue, since many shoppers had visited over half a dozen retailers and the survey was designed to administer a large battery of questions for each of 5 retailers. Indeed several market research experts consulted on this project advised us that questionnaire lengths above 20-40 minutes would cause steep losses in terms of respondent attrition, and advised us to stay within that time range since it is standard industry practice. Indeed, a recent study had found that the standard Osgood's semantic differential test, comprising just 18 questions, consumed about 45 minutes of respondent time [Ballou 2004]. Surveying the literature showed that that longest surveys were [1-2 hours?], and they were indeed rare. Further, our own pilot testing revealed that there is a psychological effect at the 1-hour mark, since many busy people juggling work and families tend to mentally block out time to answer the survey in units of 1 hour; they become more reluctant to sit down to a task if they are led to expect it will take more than an hour. Thus there was initially considerable pessimism that a survey of this scale could be pulled off successfully.

However, there was a major difference between our survey and other common marketing surveys which elicit generalities (e.g., "Most claims of product quality are true" [Bearden and Netemeyer 1999 p.351], or the items on the semantic differential test), namely that our survey was designed to elicit recall of specific events that had occurred to the respondent, in the natural logical sequence of their experiences. We avoided generalities such as questions about what "typically" happened, and focused instead on a chronological recount of what happened during the shopping experienced. This directly triggers natural cognitive processes of the kind involved in story-telling, since describing an earlier event automatically triggers recall of subsequent events. Therefore the perceived effort was significantly lower, and respondents would quickly answer numerous questions in succession, pausing at logical break points in the narrative. The survey was designed to elicit a narrative and feel like a conversation, even though most items were quite structured and not open-ended. Clearly, the process flow models constructed during the Discovery phase were key to the design of the survey, and the extra effort invested in that phase paid off substantially during survey design and execution.

A correct chronological design has a triple-benefit: not only does it reduce perceived effort, it also reduces the actual time to complete the survey, and it minimizes the introduction of model error due to irrelevant information. For example, it does not matter whether salespeople at a store have typically been very helpful in the past, if they were not helpful on the particular shopping trip for TVs. All of this indirectly improves survey completion rates as well, reducing cost.

Many other techniques were also used to reduce the cognitive load on respondents, and resolve or ameliorate the problems described earlier (Figure 27).

- We divided the survey into modules that provided natural stopping points, enabling the respondent to temporarily suspend the survey and return later to complete it. Since respondents typically visited multiple retailers, and since most questions about a particular retailer were bunched together due to the chronological narrative design mentioned above, it was natural to designate each retailer as a `module'. The survey provided clear visual indication of progress, which is another important design element. Respondents were told at the beginning how many retailers they would have to describe their experiences with, and after a battery of questions about one retailer, they were thanked, their progress was shown, and they were presented with the beginning of the next retailer. Through pilot testing we verified that respondents got a good sense of how much effort they were going to expend, and could plan accordingly whether to continue or to stop.

- Good visual design was used to reduce some of the perceived effort. Many questions were grouped together into one larger question using a few introductory sentences and then the actual questions were indented as if they were sub-questions. Only the outer question was numbered. This has the psychological effect of reducing the perceived number of questions.

- Although there were hundreds of questions in the survey, many questions were skipped for each respondent due to the underlying quasi-deterministic structure. For example, respondents who said that they did not talk to a salesperson were not asked any further questions from our Salespeople section (for that particular retailer). Or if a respondent said they had never shopped at the retailer before their TV shopping experience, we could derive the number of previous purchases at that retailer as being zero, skipping a question. Since our survey focused on actual events as opposed to generalities (see the first point above), this meant that we had far more such skip patterns than is usual in survey research. (Moreover, our ability to analyze quasi-deterministic data meant that we did not have to minimize the number of skip patterns as survey designers typically try hard to do; see Section 4.1.4.) Therefore many of the questions simply disappeared for respondents. (Over half the dataset comprised quasi-deterministically missing data.) Note that the questions that disappeared were the ones that give respondents the most trouble in typical surveys (because they are largely inappropriate even when turned into generalities that can be answered). Thus the actual effort of answering the survey was reduced substantially due to the presence of these patterns. This was yet another benefit derived from the thorough Discovery phase research.

- Not only were questions skipped in response to previous questions, but a large number of fine wording modifications (text substitutions) were programmed into the survey to carry over context from previous questions make later questions appear more natural. This, as well as a lot of wording simplifications and extensive fine tuning during the pilot tests significantly improved ease of comprehension.
- A number of standard techniques described by Dillman [2000] and others [Krosnick 1999, Friedman and Amoo 1999, Schwarz 1999] were used to optimize the design of item scales and minimize recall error and effort. Examples include:
    - the choice of 5 points for Likert-scaled question rather than 3 or 7, using agreement scales whenever possible (Strongly Agree to Strongly Disagree) rather than evaluations,
    - the use of exception responses ("Don't remember/Don't know") to avoid misuse of scale midpoints
    - the design of nonlinear frequency scales to match common memory patterns. E.g., We used weeks and months when asking about how often flyers were received ("About once a week or more often", "Two to three times per month," "Once a month," "One or two times," "Never"), whereas we used a visits when asking about how often they purchased something on visits to a store ("On almost all visits", "On many visits,", "On some visits," "On a few visits," "Never").
- Unlike common marketing scales designed to `measure a construct' through numerous items (e.g., all the facets of "service quality"), our survey was designed to elicit factual data about past events. For the most part, each item stood on its own, and did not represent an abstract construct. E.g., the respondent either read or did not read an advertising flyer; there was no separate `construct' that influenced them to visit or not-visit the store, and the item itself was the causal variable of interest, not a latent factor. Thus there was a minimum of items relating to abstract constructs or evaluations in the survey. Since the respondent's task was simple recall instead of judgment formation,43 less effort was needed.44

[43] We did have a battery of "retailer perceptions" questions that required judgment. In the end it turned out that the responses to these questions were most noisy and least useful for causal modeling purposes.

[44] Since the focus was on simple reporting of events, the analysis of instrument reliability and validity changed character. For example, evaluating Cronbach's Alpha became moot, since there wasn't a scale comprising multiple items measuring a single construct. Instead we had to rely far more heavily on ThinkAloud exercises and interviews during pilot tests to ensure that we were eliciting the data that we intended to elicit, and thus ensure validity.

- Rigorous pilot testing was used to assess and improve the quality of the survey instrument. The initial versions of the survey were tested with informally recruited recent shoppers. The production versions of the survey were conducted at a test facility in Chicago with a representative sample of shoppers from the region. About 100 recent TV purchasers were recruited using random digit dialing, utilizing the same selection criteria and screener that would be subsequently used for the actual data collection. A $100 incentive was offered to most participants. These people were brought to the facility where they filled in the survey online just as they would do at home. However, researchers watched as the respondents filled in the survey, and prompted the respondents to think aloud. Where necessary, the observers asked additional questions to clarify the respondent's comments during ThinkAloud, and to cross-check their understanding of the question's intent. The sessions were recorded, and the observers also took extensive notes. After completing the survey, respondents were interviewed for about half an hour regarding their perceptions of the survey, especially to assess perceptions of fatigue, difficulty, and how accurately the questionnaire captured their actual experience. The latter was cross-validated by obtaining a completely free-form narrative of their entire shopping experience. Their sensitivity to the level of incentive was also assessed, and we found that smaller incentives ($50 or lower) would work just as well. Since the testing was conducted in 4 iterations spanning several weeks, many survey questions were improved and reprogrammed into the online administration tool on the fly, enabling us to test the corrected version on subsequent respondents. Not only were the survey questions evaluated in this manner, but the data collected from the 100 participants was also analyzed on the fly to determine the effectiveness of the questions (e.g., whether the responses to any item were skewed towards one end of the scale, whether there was adequate variability for modeling, etc.). Thus a good sense of reliability and validity of the instrument was obtained, and at the same time a rough sense of correctness of the underlying model was also obtained during these pilot tests.

As a result of all these efforts we achieved excellent results in terms of survey completion, despite the length of the survey. Most respondents completed the survey between one and two hours, depending on the number of retailers they had visited.

One significant conclusion from our research is that survey lengths of 2 hours or longer should no longer be considered beyond the pale of normal data collection efforts fearing subject fatigue and high dropout rates. The combination of a solid ethnographic basis for design, psychometric fine tuning, and good incentive and survey strategy design will yield good data at high completion rates.

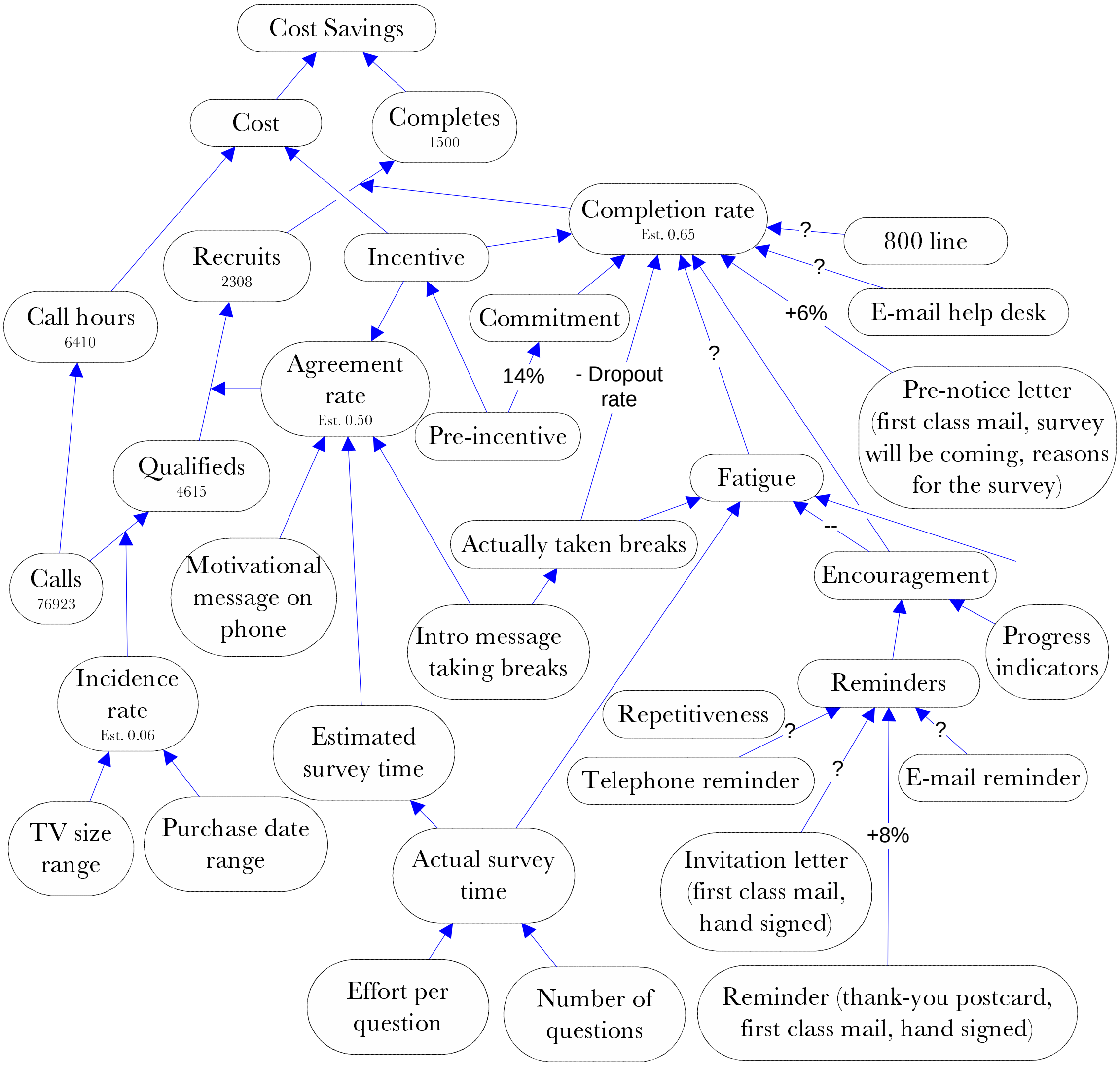


Figure 27 Some factors that affect survey completion rates and costs

4.2.3 Maximizing completion rates

Because of the significant concern about the unusual survey length and feared loss of respondents, significant attention was devoted to optimizing the survey administration strategy in order to maximize survey completion rates. This became even more critical when the incidence of qualified recruits (TV purchasers) in the population was found to be as low as 6%, which dramatically raised the cost of recruiting. Further, while some of our market research consultants suggested 50% completion rates as a satisfactory target based on industry practice in marketing, we wished to avoid the resulting concerns about bias, and preferred to target the 70-80% completion rates more typical of medical research or government surveys. So for several reasons it became extremely important to motivate and ensure that recruited respondents completed their surveys.

- Incentive design: After respondents had been successfully qualified during the telephone interview, they were offered a $50 incentive that would be given to them upon completion of the questionnaire. The appropriateness of the amount was assessed via experiments and interviews during the pilot tests of the instrument; a surprising proportion of respondents were willing to answer the survey for free; almost everyone was willing to do so for $100; $50 appeared to satisfy almost everyone. Furthermore, the initial letter of invitation included a surprise pre-incentive of 5 dollars, as a token of thanks for agreeing to participate, since this technique has been shown to be effective in increasing completion rates [Dillman 2000]. While the recruited respondents could have kept the money and omitted further participation, few did so; the success of the method has been attributed to several reasons including the change in nature of the transaction from an economic one (pay for work) to one of trust, as well as a guilt factor.

- Initial invitation: Those respondents who had an e-mail address (by far the majority) were sent an invitation message with a link to the survey immediately after the recruiting phone call. All respondents were sent a paper letter containing the invitation to the survey as well as their pre-incentive. First Class mail was used since a significant loss of participation is attributed in the industry to letters discarded unopened by respondents who associate a "Presorted" stamp with junk mail.

- Reminders: All recruited respondents who left the survey unfinished after partially completing it were sent an e-mail reminder to finish the survey. Further lack of response resulted in a series of telephone reminders, and the persistence calls resulted in significantly raising completion rates.

- 800 number support: Respondents were given an 800 number to call for help if they encountered difficulties during the survey. Most calls were either the result of unfamiliarity with using a Web browser, or from attempting to go back to an earlier stage of the survey and change their previously entered responses. While this was permitted, since it is natural to expect that the respondent's memory may as they proceed through the survey, the logical dependencies of later questions on responses to earlier questions had the potential data loss on intervening questions, and thus required some assistance.

- Progress tracking: The survey was divided into six sections: one introductory section, followed by one section for each of the five retailer types, and respondents were given an indication of how far they had progressed through the entire questionnaire, along with encouragement to proceed. This sense of perspective appeared to somewhat reduce subject fatigue; further, it provided some respondents with a logical place to stop for a break (between retailers).

- Multiple channels: When the inhibitor appeared to be difficulty of access to the Internet, surveys were administered over the phone, with the recruiter reading aloud the questions on the online survey, and filling it in based on the respondent's answers. While the telephone interviewers were trained at administering voice surveys, this does introduce a new source of possible errors and bias, in the 20% of interviews that were so conducted. However examination of the dataset during subsequent analysis did not reveal any distinction between the two groups of respondents.

- Response analysis: After we began administering the surveys and the initial datasets began to come in, we analyzed the data to find out whether there were any particular points at which significant attrition occurred. We did not find any such drop-off points, not even the breakpoints between the retailers, other than the initial welcome page. We speculate that the drop-off at the welcome page appears to be the result of people clicking on the link in their invitation e-mails out of curiosity, without being ready to fill in the survey.

[Review of drop off rates we found at various points, and completion rates. Also Check instrument fielding plan document.]

4.2.4 Adapting the quantitative methodology: some design tradeoffs

Given the cost of executing our data collection methodology[45], the question naturally arises as to what compromises can be made, trading off `gold standard' data quality for lower cost, yet producing reasonably acceptable model validity.

Clearly, such a compromise is determined in part by the subjective level of trust placed in each methodology by users of the model. It is common in the market research field to feel good about `large samples' (thousands of respondents) even when the effect sizes measured turn out to be tiny, and to feel that small sample sizes (tens of respondents) are `merely qualitative' findings, even when large effect sizes are observed, 95% significance levels are achieved, and fully representative sampling methods have been used. When we discussed sample sizes with our client, they were initially quick to tell us that 1500 respondents would be woefully adequate for our study because in their own studies they have found the need for sample sizes on the order of 150000 respondents to detect significant differences between control groups and treatment groups. It turned out that they were studying the effects of different flyer designs on sales---and as we subsequently found in our own study, the effects of flyer design are extremely slim indeed, so it is no surprise that such large sample sizes were required to detect such small effects. In contrast, since our study contained detailed casual chains, we were able to detect strong *local* effects (i.e., effects on variables adjacent in the causal chains), even when the net effect on a distant variable (the purchase outcome) was tiny. So most of the links in the model could be established at high significance levels with small sample sizes (a few hundred respondents). Given the misconceptions in the field about what constitutes good "quantitative" research versus "merely qualitative" research, any methodological decision is bound to be a partly political issue. Many managers take decisions with little or no data, and consider focus group data to be perfectly adequate; others want no less than random digit dialing before they accept any finding from a study, and will not tolerate, say, Internet panel data. The corresponding cost swings are enormous.

Our review of the sampling field gave us no reason to suspect the quality of data obtained via pre-recruited online panels except for the process of sample selection, and this process was generally so murky that we could not make any judgment about the

[45] Random digit dialing to get 1500 respondents at a 6% population incidence rate required over a million phone calls and cost over half a million dollars.

representativeness of samples obtained from such panels. This uncertainty is generally quite troubling in situations when the conclusions of the study may be strongly challenged. For example, we concluded that the brand of the product or the design of store layout has little or no causal influence on sales, despite apparent correlations. There are whole communities within the marketing profession (and at our client) who would hotly dispute and reject these contentions (often because they specialize in those subfields of marketing), and it is critical that no one can cast aspersions on the validity of the data. When such situations are anticipated, the cost of RDD might be warranted. When the situation is not so critical, online panels could provide adequate validity---after all, it is hard to construct a concrete theory about which variables in our model would be biased by the self-selection process that is inherent in online recruiting. And when such biases are identified, only a small portion of the model is typically affected.

A much better way to save cost is by controlling population incidence. When only 6% of the households reached qualify (because people do not buy high-end TVs very frequently), cost are much higher than when 90% of households qualify (e.g., most people buy groceries very frequently). High incidence rates bring costs down so much that RDD becomes an easy choice. While the product or service under study may be fixed, there are many product or respondent attributes which affect incidence rate, e.g., the sizes of TVs that qualify as “high-end” or the recency of the purchase, and these attributes can be modified to control costs. The corresponding tradeoff is usually increased noise in the model due to increased heterogeneity or increased recall-error.

The complexity of the model can also be a factor. For example, the presence of a large number of quasi-deterministic patterns or other nonlinearities results in increased analysis effort and time. However, this is only a temporary issue, until standardized analysis software becomes widely available.

4.3 Data analysis

4.3.1 Data cleaning and sample size

Of the approximately 1900 recruited respondents who at least accessed the initial page of the survey, 1541 completed the entire survey. These cases were analyzed for several kinds of errors and inconsistencies that indicated bad data, e.g., whether the recency of TV purchase that had been initially reported by the respondent at the time they were recruited and screened matched what they now reported in the online survey, or whether respondents were `Christmas-treeing’ (clicking answers without thinking them through). As a result of this cleanup, 1504 valid and complete cases were retained and used for analysis.

While all these respondents provided responses to some of the critical questions in the survey, it should be noted that the numbers of responses to other questions in the survey were significantly smaller, because of the quasi-deterministic patterns underlying the data, thus reducing the effective sample size (for some kinds of analysis). For example, only 48% of all shoppers visited Circuit City, and thus the sample size for store-experience or salespeople related questions about Circuit City is 727 instead of 1504. The lower sample size affects `local’ analysis within the affected section, e.g., analysis of the relationships between Circuit City's in-store variables, but does not affect `global' analysis because the quasi-deterministic structure captures relationships at the global level. E.g., if the respondent didn't shop at Circuit City, they didn't buy from Circuit City, and this pattern is true for *all* shoppers, not just the ones that visited Circuit City stores; thus the effective global sample size is 1504.[46]

The dataset obtained from the Confirmit survey administration system was in SPSS format and comprised two tables: about 477 retailer-independent variables in one table, and about 262 retailer-specific variables repeated over 1-5 retailers (as described in Section 4.1.5) in the second table. Since the latter table had one retailer per row (i.e., there were 1-5 rows per respondent), and because our analysis had to treat each retailer-specific item as a separate variable, this table was converted using a long-to-short transformation, thus creating 1310 variables that were then appended to the variables from the first table, to create a total of approximately 1700 variables in the dataset. Of these variables, about two-thirds were nominal (including binary variables), about a quarter were ordinal (Likert-scaled), and the rest were continuous-scaled. There was also a small set of free-text variables from open-ended questions that were discarded after descriptive analysis.

4.3.1.1 Coding of missing data

As discussed earlier, the primary source of `missing data' was the quasi-deterministic relationships, i.e., the data was not really `missing', it was logically inapplicable given the respondent's particular situation. Therefore large portions of the dataset contained missing data, whose causes were traced via the instrument and the qualitative model to the response on a prior question. For example, a respondent who said they had only visited a retailer online, but not a store, had missing data on all the Store Experience and Salesperson questions. These questions appeared within the dataset as having blank or `System-missing' values. The analysis of these variables required either the Switched-product method (Section 4.1.4.3) or the Reconstitution method (Section 4.1.4.5) depending on whether the missing cells of the variable could or could not be logically reconstituted. Since the statistical software packages (SPSS, R, etc.) could not tolerate system-missing data for anything other than basic descriptive analyses, the empty cells had to be recoded. In the case of the Switched-Product approach, since we coded the switch variable as 0 whenever the switch was "off", the product of the switch and the switched-variable was coded as 0 corresponding to the system-missing values of the original variable. In the Reconstitution approach, the system-missing values were replaced by the value that was deduced from the logic of the relationship (e.g., if the *Number of flyers received* was 0, the missing values of *Number of flyers looked through* were reconstituted as 0).

The traditional forms of `missing data' were present as well. Most questions in the survey avoided forcing an answer from the respondent, since we preferred to not get an

[46] The process of dummying the switched variable by the control variable (Section 4.1.4.3) in effect restores the effective sample size to the higher value associated with the control variable, for global patterns. However, parameters associated solely with a switched variable (i.e., not dummied by the control variable) must be computed using the lower sample size.

answer rather than get an erroneous and noisy answer. In these instances, a response of "Don't know/Don't remember" was explicitly coded as 97. During analysis, these variables were treated using standard practices for handling missing values, unlike the quasi-deterministic form of missing values (Section 4.1.4).

As described in Section 4.1.4.7, "Didn't think about it" and "Didn't care" responses really reflect quasi-deterministic relationships, not traditional missing values like "Don't remember". They were given unique codes, 98 and 99, in the survey, e.g.:

Please think back to what you thought about the following statements before you first visited Best Buy for TVs. If you did not think about a statement at that time, please select the `Didn't think about it' option.

	Strongly Agree	*Somewhat Agree*	*Neither Agree nor Disagree*	*Somewhat disagree*	*Strongly disagree*	*Didn't think about it*	*Don't remember/ Don't know*
I expected that the retailer would have high quality TVs	5	4	3	2	1	98	97

… At the time that you visited the store, how did you feel about.

	Excellent	*Very good*	*Good*	*Average*	*Poor*	*Didn't care*	*Don't Remember/ Don't know*
The looks of that TV?	5	4	3	2	1	99	97

(The numbers are the data codes for each response; they were not actually displayed on the survey.)

During analysis, such survey questions were recoded into two variables each, the binary switch variable, and the Likert-scaled perception variable. The former had no missing data by virtue of its definition; the missing data cells on the latter were recoded 0 as explained above for the switch-product technique.

4.3.2 Descriptive and correlational analyses

A number of exploratory procedures and tests were conducted in SPSS to check the data quality and obtain a basic understanding of the underlying consumer choice processes, e.g., what proportion of shoppers visited each retailer and how often, the order in which order they visited retailers, what circumstances were perceived by them as triggers to purchase TVs, the average prices of TVs considered and purchased, the performance of salespeople at each retailer, the market share of each retailer, etc. Since the majority of variables were binary or Likert-scaled, a series of nonparametric tests including the Mann-Whitney and the Friedman tests were utilized to compare retailers on each attribute.[47] These comparisons were usually consistent with common knowledge about the retailers, e.g., Best Buy was more likely to come to mind as an electronics retailer than Wal-Mart; Wal-Mart was perceived to lead the other retailers on price; and so on.

[47] The presence of quasi-deterministic patterns introduces unique problems in terms of selecting the appropriate nonparametric test. If a shopper did not visit Best Buy, there is no data available about the shopper's perception of Best Buy on in-store attributes, and the latter is not ordinary "missing data", it is quasi-deterministically missing based on *Visited-Best-Buy.* Since some shoppers visit Best Buy, some visit Circuit City, others visit both or neither, how do we compare two variables such as *Helpfulness of salespeople at Retailer A* and *Helpfulness of salespeople at Retailer B* whose missingness is deterministically affected by *Visited Retailer A* and *Visited Retailer B* respectively? Clearly the missing cases cannot be argued to be independent of the non-missing ones; the shoppers who chose not to visit *Retailer A* may have done so because they believed the salespeople would not be helpful. We were not able to find a statistical test that perfectly matched our needs, and in the end we used a combination of the Mann-Whitney test for two independent samples, and the Friedman test for dependent samples, looking for consistency between the two tests. Our reasoning was as follows.

The Wilcoxon matched-pairs and Friedman tests are applicable to the subset of people who visit two retailers. The interpretation of a significant result is that people who visit both stores do not provide equal ratings to the two stores; one store has significantly better ratings than the other---among those who visit both stores. Of course these tests cannot be used to make inferences about the people who visited only one of the two stores, and since the tests strip missing values, such people are automatically ignored in the output. Deterministic missing values do not appear to be an issue, since the sample is used to make inferences about the (sub) population that visits both stores, not to the larger population that visits either or neither. The mean ranks in the Freedman test have a particularly useful interpretation: they range from (1.5, 1.5) when the two groups are similar, to (1, 2) or (2, 1) when one of the groups is ranked higher than the other by every respondent. The difference between the two mean ranks, e.g., 0.4 in (1.3 1.7), is the proportion of the population by which the second group has a lead over the first group.

The Mann-Whitney test for two independent samples appears to be suitable to comparing people who visited either of the two retailers, even though the two groups are not entirely independent since they include people who visited both retailers and thus provided two ratings. One justification is that the number of people who visited both is usually small compared to the number that visited either, so any bias from the double-ratings should be small. Secondly, if we treat one group as a regular sample, look into the second group and ask why the answers of the overlapping people should be biased, there does not appear to be any source of bias other than the possibility that those people have already provided one rating and may shift their second rating to be relative to the first one, rather than an absolute answer. This is unlikely given the structure of the survey--the questions are separated quite a bit in time--and the shifts are not likely to be large, so again any bias should be small.

However, the deterministic missing values may be an issue: we already know a-priori that the two groups come from different subpopulations: the ratings for one store are obtained only from the people who visited that retailer. The tests ignore that fact, and test whether the groups come from the same population in terms of ratings on the given variable, e.g., sales experience. So a significant result leads to the conclusion that the two groups could not have come from the same population (otherwise the offers would have been equally good). The selection-basis for the two groups (the fact that the groups are selected based on visits to each retailer) may be one explanation for the difference, e.g., perhaps the people who rated Sears service poorly did so because most of them didn't visit Wal-Mart and see how bad the service at Wal-Mart is; if more people in the population visit Wal-Mart, perhaps they would give higher ratings to Sears. Such explanations based on the deterministic missing values do not invalidate the use of the tests, since the tests merely certify that the two samples do not come from the same population with respect to service, and do not say anything about the cause of the difference in ratings.

So the main point to remember in the use of the Mann-Whitney test is to use a more conservative cut-off for the significance level than presented in the output, because of the presence of some overlapping cases

There were a few surprises, however. While we expected Best Buy to lead the other retailers on awareness (Figure 28), it turned out that almost all major retailers had about 75% awareness among the general population of TV buyers. However, this stood in stark contrast to the disparity in visit rates. While 75% of those who knew Best Buy sold TVs chose to visit the retailer, barely 43% of those who knew that Sears sold TVs ended up visiting Sears. Similar disparities were observed in terms of conversion rates: Wal-Mart was the most effective retailer, converting nearly half of its visitors into purchasers, whereas Sears and Circuit City converting about a quarter.

We also conducted standard exploratory examinations of bivariate correlation tables across most of the important variables. Again this mostly yielded a confirmation that variables that were expected to relate to the final choice of retailer were indeed correlated with that variable. The extent of correlation ranged from none to moderate, with very few variables at the higher end. This served to give us a rough feel for the relationships between variables in the model that would be subsequently constructed. However, this exploratory analysis was *not* used to guide model structuring, since we would subsequently use the much more robust causal discovery methods instead. These methods use bivariate correlations in an extremely conservative manner for variable elimination, but not for guiding imposition of structure with substantive knowledge.

Thus, descriptive and correlational analyses served to (a) construct an understanding of the typical shopper's situational context and the typical performance levels of each of the retailers and (b) generate a few provocative questions and hypotheses that causal modeling would have to explain. Beyond these two uses, descriptive analysis played little role in our analysis. It is worth noting that this is in sharp contrast to methods used in the majority of market research studies (e.g., Forrester [Kolko et al. 2003], Electronics Report, Retail Forward [2004], IBM IBV [Ballou et al. 2005], etc.], which rely primarily on descriptive data to `support' almost all their inferences and theories about the marketplace. Our view is that patterns observed during these analysis can at best suggest hypotheses (e.g., that people may have bought at a given retailer mainly because of the

between the two groups. There is no precise measure of this (theoretically, one would use a lower df in the test, reducing it by the number of overlapping cases.)

Another approach to handle the overlapping cases (the people who visited provided ratings for both S and W), is to randomly select one of their answers, thus randomly assigning them to one of the two groups. (Their rating for the other retailer is discarded.) While this removes the overlap between samples, it does raise the question of what probability to use in assigning a given respondent to the two groups; the ones that stand out are 0.5, and ratio-of-visit-probabilities. The latter seems preferable because the sample sizes of the non-overlapping respondents (the number of people who visited only S, and only W) are determined by the incidence of such people in the general population, and it makes sense that the overlapping respondents are split in the same proportions. This removes any concern of sample dependency. It does not remove any effects from deterministic missing values (because the non-overlapping respondents are allowed to answer only about the retailer that they visited, and thus Visit becomes the immediate candidate for explaining a significant result).

The Mann-Whitney test should really be replaced by the Kruskal-Wallis test + post hoc analysis, because we're doing pairwise comparisons of multiple groups. SPSS does not provide the post hoc analyses, and the K-W test itself is not useful when applied to 5 groups simultaneously. So we made do with the M-W test, and applied one correction: the cutoff p value was made more stringent than 0.05 to compensate for the increased probability of Type 1 error.

price), but all too often, robust causal analysis contradicts those hypotheses (e.g., picture quality was correlated with price and purchase, and was a common cause of both; therefore manipulating price does not have much effect on purchase, and retailer would be wrongly advised by these descriptive analyses to focus on price). Thus these analyses play only a limited role in a methodology that has a solid causal foundation. Indeed, in our methodology descriptive analysis becomes more useful *after* causal modeling has been completed, in order to generate visualizations that pictorially render the workings of the marketplace, the consumer's behavior, and their choice processes, and thus to help explain the findings from the causal model.

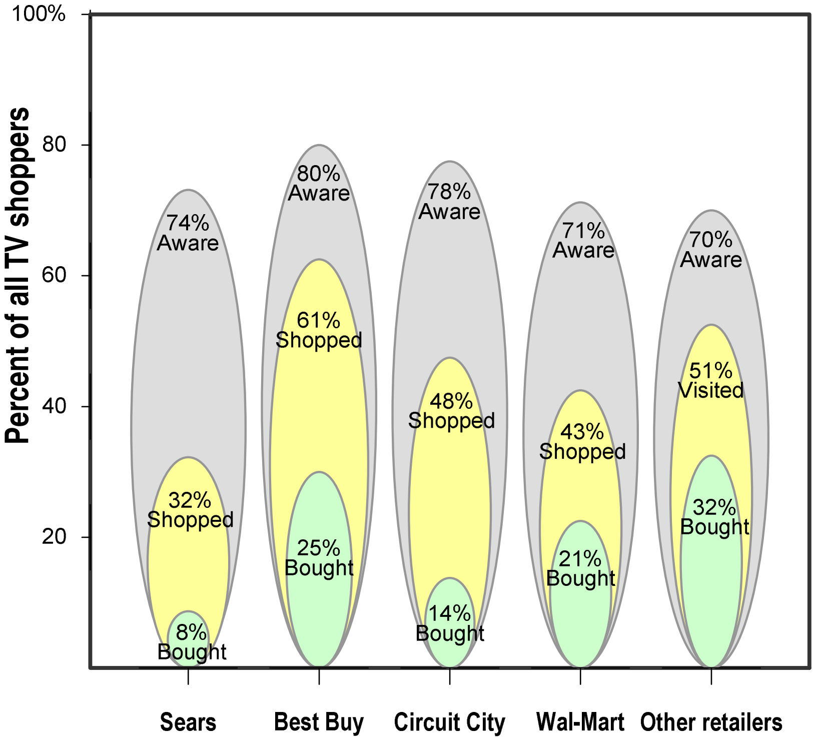


Figure 28 Comparison of awareness, visits, and purchases

4.3.3 Causal modeling and parameter estimation

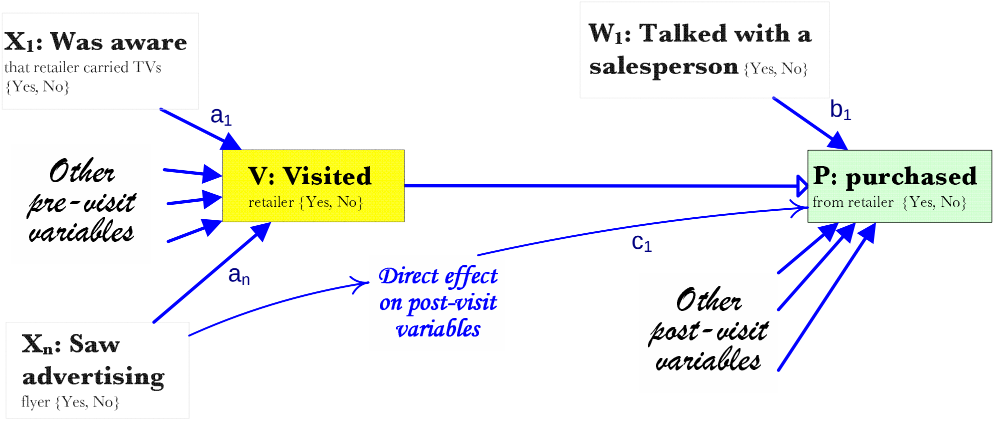
The resulting dataset was then analyzed using the causal modeling methodology described at length in Section 4.1.2. Utilizing the binary version of our outcome variable (Purchase vs. non-purchase from Retailer A), a correlation check revealed that a majority of variables had small or trivial correlations with the final outcome, though they had substantial correlations with causally adjacent variables. In other words, although many variables appeared to play a significant role at a given point in the purchasing experience, their net effect on the purchase outcome was often quite reduced due the long causal chains and large number of intervening variables between the given variable and the outcome. (Note that since many questions were asked once for each of the 5 retailers, it was predictable that only a small number of variables describing shoppers' experiences at retailer B, C, D, and E had much effect on purchases at retailer A; it was more typical for variables describing any given retailer to significantly influence purchases at that retailer. As long as we focused our analysis on outcomes at Retailer A, many variables describing retailers B through E became relatively weak in influence, and only a few variables continued to display cross-interactions between retailers.) We therefore used the correlation-check as a heuristic to screen out variables and reduce our focus to a few dozen important variables.[48] Even though a low marginal correlation with the outcome does not imply that the variable is causally unimportant, given the huge effort required to do the causal analysis (using the rudimentary algorithmic tools that we had begun developing at the time), we felt that it was worth using the screening heuristic at the small risk of dropping some important variables.[49]

The model was split into two parts, pre- and post-visit (Figure 29; also see the discussion in Section 4.1.6). Although *Visited-Retailer* is a quasi-deterministic switch variable and could have been treated using the techniques described in Section 4.1.4, it controlled numerous other variables (about half the model) and caused their data to be quasi-deterministically missing. Also, early versions of our model showed that only a few pre-visit variables had post-visit influence on the purchase decision. In other words, it was possible to break the model at the *Visited-Retailer* variable and roughly assume independence (conditional on *Visited-Retailer* of course) between variables on either side, without much loss of accuracy in estimating the total causal effect of each pre-visit variable. (The effect of post-visit variables on purchase was unaffected by such a break of course.) This effectively resulted in two sub-models: one submodel explaining the effect of all pre-visit variables on the decision to visit or not visit the retailer, and the other sub-model explaining how post-visit variables affected the purchase decisions. The two sub-models could be combined by simple multiplication because of the quasi-deterministic pattern (if the shopper did not visit, they did not purchase) and the approximate conditional independence. This brings both conceptual clarity as well as analytical simplification to the overall model of purchasing behavior.[50] Furthermore, this structure has important implications for retailers; see Section 4.4.2.

[48] A correlation magnitude of 0.1 was used as the cutoff. Variables with extremely low sample sizes (due to missing data, usually from chains of quasi-deterministic patterns) were also eliminated because their effect sizes could be theoretically predicted to be small---recall that their effects are weighted by the mean values of their switching variables. Some variables thus eliminated were restored because of their theoretical significance, i.e., it was important to explicitly demonstrate that their causal effects were small and to explain why.

[49] The use of a low marginal correlation as a heuristic to ignore causal influence is a not-so-disguised use of the stability (or faithfulness) assumption---if an explanatory variable did indeed have significant causal influence on the outcome variable, some accidental cancellation much have occurred across parallel paths in order to produce the low correlation. While we have argued earlier in this paper that stability violations are common when conditioning on effect variables, since we're only using marginal correlations in this heuristic, we argue that the chance of a cancellation must be significantly lower. However, the extensive presence of nonlinearities such as quasi-deterministic patterns gives us some reason for discomfort with this heuristic, and a more thorough empirical investigation of accidental cancellations under such conditions would be desirable to justify the heuristic.

[50] Because the post-visit model now needed data only from shoppers who had visited the retailer, all the cases corresponding to non-visitors could be discarded, and instantly a large number of quasi-deterministically missing variables could be treated as regular variables without missing data. Of course there were many post-visit variables which caused data to go quasi-deterministically missing, but the number of such variables was significantly reduced.


X1: Was aware
that retailer carried TVs
{Yes, No}
a1
Other pre-visit variables
V: Visited
retailer {Yes, No}
an
Xn: Saw advertising
flyer {Yes, No}
Direct effect on post-visit variables
W1: Talked with a salesperson {Yes, No}
b1
P: purchased
from retailer {Yes, No}
c1
Other post-visit variables

$$
\begin{aligned}
P &= V \times \left(b_1 W_1 + \ldots + b_m W_m + c_1 X_n\right) \\
&= \left(a_1 X_1 + \ldots + a_n X_n\right) \times \left(b_1 W_1 + \ldots + b_m W_m + c_1 X_n\right) \\
&\approx \left(a_1 X_1 + \ldots + a_n X_n\right) \times \left(b_1 W_1 + \ldots + b_m W_m\right)
\end{aligned}
$$

Figure 29 The multiplicative approximation into two submodels

To build each sub-model, the variables in the submodel were processed through the causality detection algorithms summarized in Figure 19. Causal assumptions were introduced as needed; the vast majority of these were of the "B Not-Cause-Of A" format and judged to be quite weak (i.e., quite defensible). A small fraction of the assumptions were of the "DualCause" format, where the two variables could simultaneously be causes of each other. Because there were only a small number of such circular relationships in proportion to the size of the model, for the most part we did not run into SEM estimation problems.

The autogenerated SEM was then estimated using the Mplus package. As described in the methodology section, we routinely obtained excellent fit on the standard indices (Chi-Square p=0.99, RMSEA = 1, CFI/TLI = 1, etc.). Nonetheless, modification indices were examined and used to flag errors made by the causality detection algorithms as described in Section 4.1.2.5; the errors were fixed by restoring incorrectly deleted links and rerunning the entire model-generation process.

The default SEM was generated in a form that treated the outcome as a linear variable, even though the data was binary. This approximation worked quite well, yielding an R-square of 0.65. Declaring the outcome as a "Categorical" variable in Mplus initially

In general, this technique of splitting the model into multiple parts is not possible because the resulting submodels are not multiplicative; variables on the left of the split are not independent of (and can causally affect) variables to the right of the split. To compute the total causal effect of a variable on the left of the split on the outcome variable on the right of the split, causal paths passing through the switch-variable as well as paths not passing through the switch variable must be added up. This can be an extremely complicated procedure if the two submodels are built separately, but becomes straightforward using the methods described in 4.1.4.

triggered convergence problems[51] but eventually produced a working model with an R-square improvement of around 10%. Note that declaring a dependent variable as categorical in Mplus results in the estimation of a Probit model for that variable [Muthen 1998-2004].

4.3.4 Identifying the critical factors

Total effects were calculated by the Mplus software, and corrections were made to compute the total effects of quasi-deterministically switched variables (Section 4.1.4.4). The results were examined to identify the most important variables affecting purchase. The top ten variables were called out as the most critical factors for our client.

The standardized total effect of a variable on the purchase outcome was used as a measure of importance of the variable. We were well aware of the meaningless nature of this metric; see the discussion in Section 4.1.3. As described there, we believe the right way to do so is by estimating the `return on intervention', i.e., the gain in sales, measured in dollars, divided by the cost of making an intervention, also measured in dollars. However, the procedure we recommend requires that we obtain or guess at the cost per unit intervention for each variable. In our study we were not able to obtain those estimates due to constraints on time and availability of our client. So we fell back on the standardized total effect as the criterion for comparison, unsatisfactory though it is.

Note that the factors are generally not independent of each other. In fact, one factor often causally affects several other factors. So, for simplicity we ranked them assuming we change only one factor at a time. The Top Ten ranking served primarily a psychological purpose: our client needed something short and sweet that they could carry in their minds, and a full-blown causal model was not appropriate for that purpose. However when computing the effect of a group of interventions made simultaneously, the actual causal model would have to be used, since improvements from the Top Ten factors are not always additive.

If we consider just three of the Top Ten factors, whose effects do happen to be additive, we get an impressive finding: by

1. Getting visitors to seriously consider at least one TV at the retailer
2. Creating the feeling of "getting a deal", and
3. Targeting shoppers who care about better credit and financing terms

our client can double their sales by intervening on these simultaneously.

[51] The numerical problems appear to arise from two sources: Mplus switches the estimation method from maximum likelihood to least squares and uses Theta parametrization when the the dependent variable is categorical, and appears to have bugs that are triggered by the presence of categorical explanatory variables in the model. We used Mplus 3.1; more recent versions appear to have at least partially addressed some of these problems.

4.3.5 Identifiability

Traditional analysis techniques used in confirmatory structural equation modeling to determine the identifiability of structural parameters and total effects turned out to have little use in our study, despite our making a concerted effort to use these techniques. The enormous size of the model and the presence of numerous nonlinearities (Section 4.3.6) precluded either manual or automated identifiability analysis---in situations when we *had* a hypothesized structural model to analyze. The qualitative causal model manually specifies only first-class causal links, and the omission of other links, which was necessary for reasons of scale, prevents identifiability analysis on that particular model. As our quantitative methodology became more reliant on semi-automated discovery of causal structure and auto-generation and testing of several hypothetical structural models, it became quite impractical to perform any a priori identifiability analysis at all. Indeed, the causal discovery algorithms dynamically decide the presence or absence of structural links and their directionality, which affects identification. The problem was aggravated with the introduction of feedback loops (causal links going in both directions between two variables, or directed circular paths chaining over a number of variables), which turned out to be quite plausible for several links. It turned out to be methodologically simpler to just export the auto-generated structural equation model into Mplus and let the software detect identifiability problems during model estimation, typically in the form of unestimated parameters. This worked in practice because (a) the causal structure discovery algorithms were good enough to find structure that rarely had identifiability issues and (b) when identifiability issues did surface, the Mplus software output provided enough clues to diagnose the source (e.g., feedback loops that involved a latent variable) relatively easily, and fix it via a slight adjustment to the model. When manual adjustments to the model were attempted, the most useful rule of identifiability was the two-output rule [Ken Bollen (personal communication); Jarvis et al. 2003].

4.3.6 Nonlinearities in the model

It is worth noting that there were several sources of nonlinearity in the model that preempted the use of traditional analytic tools. The most distinct source of nonlinearities was of course the ubiquitous quasi-deterministic pattern; in each instance of this pattern, the switched variable could not be written as a linear function of its control variable or of its non-local causes. Even though our switched-product procedure (Section 4.1.4.3) restored a modicum of linearity to the analysis, significant care was required in the interpretation of structural parameters associated with the switched variable, and the causal modeling algorithms were also affected.

The second source of nonlinearities arose from the prevalence of a large number of categorical variables in the model. While some categorical variables such as the final outcome (the discrete choice of retailer) were inevitable, their large number throughout the model was in part due to our drive toward eliciting hard facts, e.g., specifying whether an event did or did not occur, and our avoidance of fuzzy constructs such as attitudes which typically have Likert scales that are commonly treated as continuous variables in structural equation models. Again, these were handled to a certain degree with linear tools utilizing the GLM (probit) transformation, but due to the presence of these categorical variables in many intermediate causal chains throughout the model (not just as the outcome variable) they posed considerable computational/convergence issues.

A third source of nonlinearities arose in certain parts of the model where the functional relationship between some variables was so intrinsically nonlinear in form that approximation by linear function would produce significant error. For example, the submodel describing how product attributes influence purchase is intrinsically divided into the “screening” attributes and the “feeling” attributes, and the effect of each screening attribute is multiplicative instead of additive (Section 4.4.3.1). Similarly the effect the frequency with which advertising flyers are received by shoppers on their likelihood of visiting a retailer appears to be logarithmic: as more flyers are sent, there is a diminishing improvement in response.

4.4 Findings

The detailed findings from the quantitative model, as well as the final causal model of consumer purchasing, are documented elsewhere [Kramer and Noronha 2005]. We mentioned one of our striking findings earlier: out of the hundreds of variables that we considered, interventions on just three variables would double our client’s sales. We highlight a small selection of other noteworthy findings here.

4.4.1 Contradicting common wisdom: some spurious correlations

Within the marketing community, belief in the importance of `brand’ for driving sales and customers loyalty appears to reach the level of a deep act of faith. Any attempt to challenge that belief is met with considerable resistance. On one hand, there is indeed evidence that, e.g., some well-known brands inspire strong attachment among some consumers. Yet even a casual field trip to a retail store reveals that many consumers switch brands quite readily, or do not appear to think about the brand when choosing their purchases. Given the many millions of dollars routinely invested by manufacturers and retailer to “build their brand”, there is surprisingly scant research on exactly how much return in terms of increased revenue is yielded by a dollar invested in brand-building. Of course, the difficulty lies in causally connecting an intervention (e.g., an advertising program) to the increased revenue when numerous other factors can confound the problem.

We observed a particular instance of such confounding in our study: our client pointed out that every time they advertised a product in their weekly flyer, there was an increase in sales of that product immediately after. However, they omitted to note that every time they placed a product on sale in the flyer, the product was also placed on sale within the stores, i.e., there was promotional tags attached to those products, calling attention to the lower prices. A shopper who walks into the store having never seen a flyer could still be attracted to the product because of the promotional offer. Clearly the in-store intervention confounds the effect of the flyer.

During our Discovery Phase research, we observed shoppers refer to TVs by brand name, e.g., the “32 inch Sony”, the “36 inch Toshiba”, etc. However, it was also obvious from the ThinkAloud data obtained when the shoppers compared TVs that they did not care about the brand; they were focused on other characteristics such as the size of the TV, the picture quality, and so on. On closer inspection, we came to a realization that to us was a significant insight: the shoppers were simply using the brand-name as a `handle', a way to reference an object. They needed some way to refer to one TV and then to another; this was the verbal analog of physically pointing to a TV, in order to identify it as the subject of discourse. The brand itself was meaningless except as a name for a thing; there were no preferential connotations favoring one brand versus another.[52]

While we had qualitative reasons to discount the role of brand in influencing sales, how did brand stack up in the quantitative research? Here are the highlights:

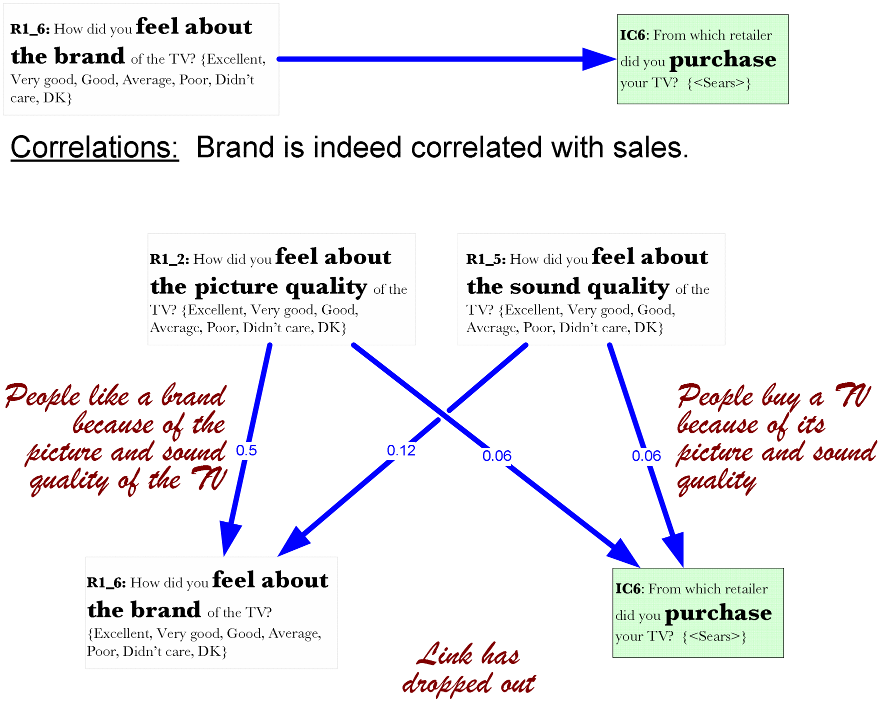
- Although 92% of visitors to a retailer say they care about brand, on the whole the model finds that their feelings about the brand of the TV they liked best at the retailer has a weak *effect* on sales.
- Brand does *correlate* with sales, but much of the correlation is due to the underlying picture and sound quality of the TVs.
- Brand does have some influence on overall *Liking for a TV*, but the net effect on *Purchase,* which is further down the causal chain, becomes quite weak.

The power of causal modeling is most obvious when alternative explanations are discovered, throwing light on why brand is correlated with sales but does not causally influence sales. Figure 30 provides that explanation. Picture and sound quality are the factors that induce a correlation between brand and purchase. People like a brand when those TVs have a good picture and sound quality. People buy TVs with good picture and sound quality. Therefore a brand that is liked becomes a brand that is purchased. You could change the brand itself to anything else; as long as the picture and sound quality are good, the product will continue to sell.

We suspect this result can be generalized well beyond our study, viz., one cannot infer that brand has an effect on purchase until one has accounted for other product attributes that might correlate with brand. Much of the popular belief in the importance of brand is merely supported by the association between brand and purchase. As our exposition illustrates, relying on this association is excusable but naïve reasoning; the correlation may be readily shown to be spurious in light of product quality and other attributes.

Apart from theoretical implications, there are important practical implications for our client. Within their stores most of the valuable wall-space was dedicated to displaying the brands of TVs that they carried. Instead they would do much better to devote that wall-space to advertising that they offer customers a really good deal on their purchases---this being one of the top ten critical factors identified by our study, and one of our client's weak points.

[52] This is an over-simplification. Many shoppers did indeed distinguish between regular brands and low-end brands in the sense that they would not purchase the low-end TVs such as Audiovox. And some shoppers considered Sony TVs to be better than the rest. With those exceptions, most shoppers treated all the regular TV brands (Toshiba, Panasonic, JVC, etc.) as indistinguishable. So, `brand' does make a small difference, but clearly does not deserve the disproportionate significance given to it by marketers.


R1_6: How did you feel about the brand of the TV? {Excellent, Very good, Good, Average, Poor, Didn't care, DK}
IC6: From which retailer did you purchase your TV? {<Sears>}
R1_2: How did you feel about the picture quality of the TV? {Excellent, Very good, Good, Average, Poor, Didn't care, DK}
R1_5: How did you feel about the sound quality of the TV? {Excellent, Very good, Good, Average, Poor, Didn't care, DK}
People like a brand because of the picture and sound quality of the TV
0.5
0.12
0.06
0.06
People buy a TV because of its picture and sound quality
R1_6: How did you feel about the brand of the TV? {Excellent, Very good, Good, Average, Poor, Didn't care, DK}
IC6: From which retailer did you purchase your TV? {<Sears>}
Link has dropped out

Correlations: Brand is indeed correlated with sales.

Causal effects: TV brand does not have much effect on sales.

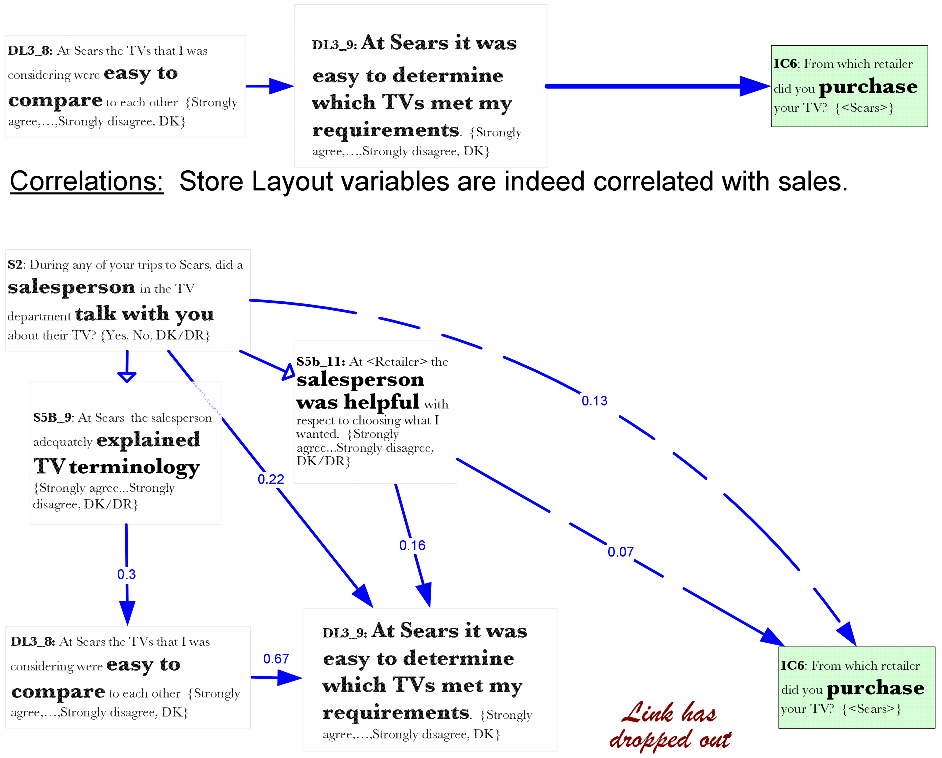
Figure 30 The causal effect of brand

A second major example of common wisdom going wrong relates to store layout: good store layout was hypothesized to drive sales. Again our quantitative study showed that the correlations do indeed support that theory (Figure 31). However, our causal model showed that there is a weak causal link at best.

The alternative explanation of the correlation turned out to be `help from salespeople'. In retrospect, the explanation seems obvious. What does a shopper do when they cannot find the kind of TV they want? Talk to a salesperson of course. Thus salespeople compensate for poor store layout, and make it easier to find the product, inducing a correlation between *talking to a salesperson* and *easy to determine which TVs met my requirements*. Salespeople also influence shoppers' perceptions and affect whether shoppers seriously consider any TV at all, which in turn has a big effect on sales. Thus there is a directly correlation between *talking to a salesperson* and *purchase*. Salespeople are thus the common-cause that induces the spurious correlation between *easy to determine TVs* and purchase.

Again, the practical impact of this finding on retailers is huge. There is a huge industry around store design and merchandizing, with retailers spending millions of dollars on frequent changes to store format. Our results imply that salespeople-related interventions may be a better investment than floor renovations.[53]

[53] However, see our caveat in Section 4.3.4 about measuring `return-on-intervention'. We do not know how much it will cost to improve salespeople's behavior; changing people is often difficult even with good


DL3_8: At Sears the TVs that I was considering were easy to compare to each other {Strongly agree,…,Strongly disagree, DK}
DL3_9: At Sears it was easy to determine which TVs met my requirements. {Strongly agree,…,Strongly disagree, DK}
IC6: From which retailer did you purchase your TV? {<Sears>}
S2: During any of your trips to Sears, did a salesperson in the TV department talk with you about their TV? {Yes, No, DK/DR}
S5b_11: At <Retailer> the salesperson was helpful with respect to choosing what I wanted. {Strongly agree...Strongly disagree, DK/DR}
S5B_9: At Sears the salesperson adequately explained TV terminology {Strongly agree...Strongly disagree, DK/DR}
0.13
0.22
0.16
0.07
0.3
0.67
DL3_8: At Sears the TVs that I was considering were easy to compare to each other {Strongly agree,…,Strongly disagree, DK}
DL3_9: At Sears it was easy to determine which TVs met my requirements. {Strongly agree,…,Strongly disagree, DK}
Link has dropped out
IC6: From which retailer did you purchase your TV? {<Sears>}

Correlations: Store Layout variables are indeed correlated with sales.

Causal effects: Store Layout variables do not have much effect on sales.
The common-cause, Salespeople's behavior, explains away the correlation.

Figure 31 The causal effect of store layout on sales

4.4.2 Implications from the structure of the model

As described in an earlier section (4.3.3), we discovered that the model could be cleaved into two submodels, pre- and post-visit, because few pre-visit variables had much *additional* influence on purchase (other than accounted for by their influence on driving visits). The two submodels are multiplicative in nature, which has important implications for retailers (Figure 32). It is obvious that doubling visits roughly doubles sales. More importantly, the return on pre-visit interventions is weakened by low in-store conversion rates and the return on in-store interventions is weakened by low visit rates. For example, consider the data for Sears as shown in the figure. Since only 32% of TV shoppers visit Sears, a good in-store intervention that improves the conversion rate of shoppers into buyers by 20% will produce only a 6% increase in market share. Furthermore, since only 26% TV shoppers who visit Sears buy from Sears, in order to increase market share by 10% using pre-Visit interventions such as Advertising, Sears must drive at least 40% additional shoppers to visit Sears. Therefore we conclude that it

training or hiring programs. If salespeople-related interventions are expensive, other variables may be considered to be better investments even if they have smaller influence on the outcome. Regardless of the question of what is the best investment, the causal model presented here makes it clear that investing in store layout improvements will not yield good results; the correlation is spurious and does not represent an underlying causal link.

is important to simultaneously make both types of interventions; it is not sufficient to focus, say on store improvements and expect to get great improvements on sales if visits have not been simultaneously raised. Put differently, pre- and post-visit interventions leverage each other and there is a multiplier effect.

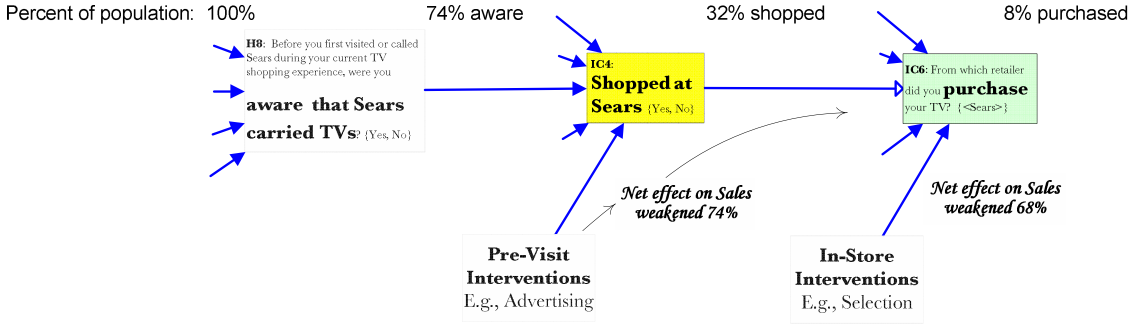

Percent of population: 100%
74% aware
32% shopped
8% purchased
H8: Before you first visited or called Sears during your current TV shopping experience, were you aware that Sears carried TVs? {Yes, No}
IC4: Shopped at Sears {Yes, No}
IC6: From which retailer did you purchase your TV? {<Sears>}
Net effect on Sales weakened 74%
Net effect on Sales weakened 68%
Pre-Visit Interventions E.g., Advertising
In-Store Interventions E.g., Selection

Figure 32 The effects of pre- and post-visit interventions

The point is made more starkly if we contrast the above numbers against the corresponding numbers for Best Buy. Since Best Buy performs better than Sears on both visits and conversion rates (Figure 28), Best Buy can get greater returns from exactly the same interventions as Sears. E.g., a $10M investment in store improvements would touch 61% of the target market at Best Buy but only 32% at Sears, thus producing nearly twice the return on investment.

4.4.3 Findings from specific sections of the model

4.4.3.1 Product characteristics

We found product characteristics did not have a straightforward linear effect on the decision to purchase. Instead we discovered that product characteristics influence the purchasing outcome via two distinct mechanisms, and we classified product attributes correspondingly as `Screening' and `Feeling' attributes (Figure 33). Screening attributes are a match/fail proposition; if the attribute matches the customer's needs, he or she will consider the TV for purchase, and not consider the attribute any further, say when comparing two TVs that qualify. For example, if a TV is too large to fit in the customer's entertainment center, it won't be considered for purchase regardless of how nice the product looks and how good a deal is offered on it. On the other hand, *Picture quality* is classified as a Feeling attribute because the shopper will compare multiple TVs on this attribute and pick the one that appears to have better picture quality.

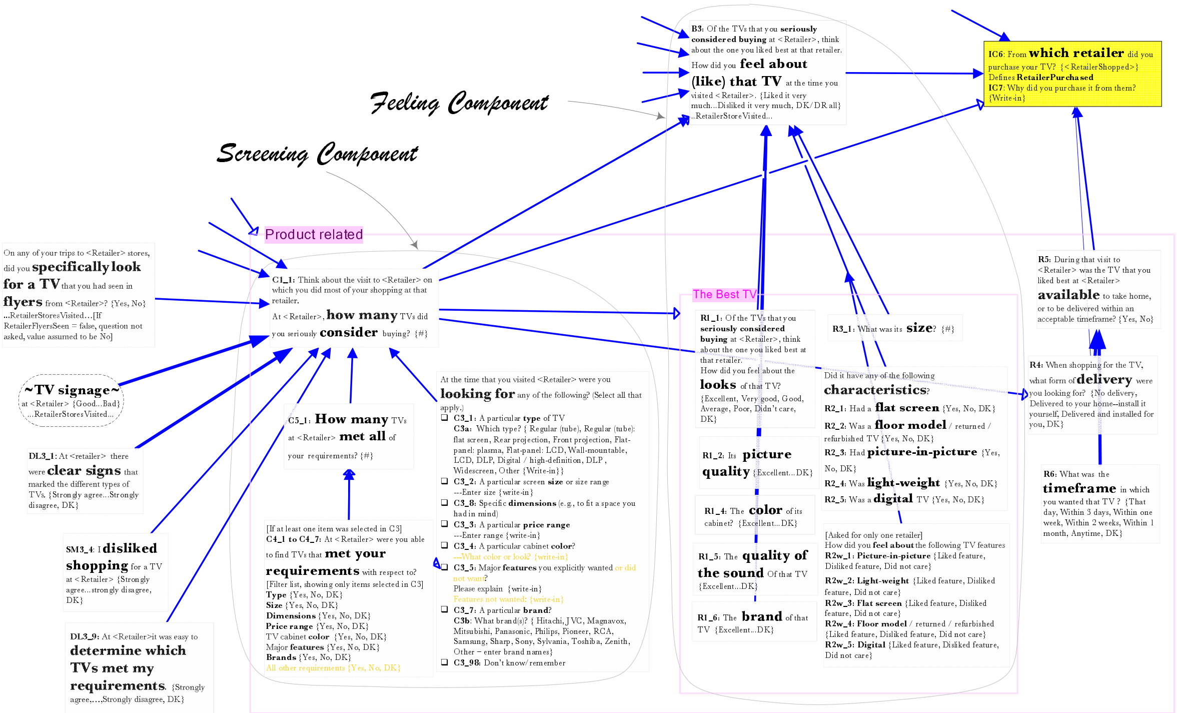

Feeling Component
Screening Component
Product related
On any of your trips to <Retailer> stores, did you specifically look for a TV that you had seen in flyers from <Retailer>? {Yes, No} ...RetailerStoresVisited…[If RetailerFlyersSeen = false, question not asked, value assumed to be No]
~TV signage~ at <Retailer> {Good...Bad} …RetailerStoresVisited…
DL3_1: At <retailer> there were clear signs that marked the different types of TVs. {Strongly agree...Strongly disagree, DK}
SM3_4: I disliked shopping for a TV at <Retailer> {Strongly agree...strongly disagree, DK}
DL3_9: At <Retailer>it was easy to determine which TVs met my requirements. {Strongly agree,…,Strongly disagree, DK}
C1_1: Think about the visit to <Retailer> on which you did most of your shopping at that retailer.
At <Retailer>, how many TVs did you seriously consider buying? {#}
C5_1: How many TVs at <Retailer> met all of your requirements? {#}
[If at least one item was selected in C3]
C4_1 to C4_7: At <Retailer> were you able to find TVs that met your requirements with respect to?
[Filter list, showing only items selected in C3]
Type {Yes, No, DK}
Size {Yes, No, DK}
Dimensions {Yes, No, DK}
Price range {Yes, No, DK}
TV cabinet color {Yes, No, DK}
Major features {Yes, No, DK}
Brands {Yes, No, DK}
All other requirements {Yes, No, DK}
At the time that you visited <Retailer> were you looking for any of the following? (Select all that apply.)
❑ C3_1: A particular type of TV
C3a: Which type? { Regular (tube), Regular (tube): flat screen, Rear projection, Front projection, Flat-panel: plasma, Flat-panel: LCD, Wall-mountable, LCD, DLP, Digital / high-definition, DLP , Widescreen, Other {Write-in}}
❑ C3_2: A particular screen size or size range
---Enter size {write-in}
❑ C3_8: Specific dimensions (e.g., to fit a space you had in mind)
❑ C3_3: A particular price range
---Enter range {write-in}
❑ C3_4: A particular cabinet color?
---What color or look? {write-in}
❑ C3_5: Major features you explicitly wanted or did not want?
Please explain {write-in}
Features not wanted: {write-in}
❑ C3_7: A particular brand?
C3b: What brand(s)? { Hitachi, JVC, Magnavox, Mitsubishi, Panasonic, Philips, Pioneer, RCA, Samsung, Sharp, Sony, Sylvania, Toshiba, Zenith, Other – enter brand names}
❑ C3_98: Don't know/ remember
B3: Of the TVs that you seriously considered buying at <Retailer>, think about the one you liked best at that retailer.
How did you feel about (like) that TV at the time you visited <Retailer>. {Liked it very much...Disliked it very much, DK/DR all} ..RetailerStoreVisited...
IC6: From which retailer did you purchase your TV? {<RetailerShopped>}
Defines RetailerPurchased
IC7: Why did you purchase it from them? {Write-in}
The Best TV
R1_1: Of the TVs that you seriously considered buying at <Retailer>, think about the one you liked best at that retailer.
How did you feel about the looks of that TV?
{Excellent, Very good, Good, Average, Poor, Didn't care, DK}
R1_2: Its picture quality {Excellent...DK}
R1_4: The color of its cabinet? {Excellent...DK}
R1_5: The quality of the sound Of that TV {Excellent...DK}
R1_6: The brand of that TV {Excellent...DK}
R3_1: What was its size? {#}
Did it have any of the following characteristics?
R2_1: Had a flat screen {Yes, No, DK}
R2_2: Was a floor model / returned / refurbished TV {Yes, No, DK}
R2_3: Had picture-in-picture {Yes, No, DK}
R2_4: Was light-weight {Yes, No, DK}
R2_5: Was a digital TV {Yes, No, DK}
[Asked for only one retailer]
How did you feel about the following TV features
R2w_1: Picture-in-picture {Liked feature, Disliked feature, Did not care}
R2w_2: Light-weight {Liked feature, Disliked feature, Did not care}
R2w_3: Flat screen {Liked feature, Disliked feature, Did not care}
R2w_4: Floor model / returned / refurbished {Liked feature, Disliked feature, Did not care}
R2w_5: Digital {Liked feature, Disliked feature, Did not care}
R5: During that visit to <Retailer> was the TV that you liked best at <Retailer> available to take home, or to be delivered within an acceptable timeframe? {Yes, No}
R4: When shopping for the TV, what form of delivery were you looking for? {No delivery, Delivered to your home--install it yourself, Delivered and installed for you, DK}
R6: What was the timeframe in which you wanted that TV ? {That day, Within 3 days, Within one week, Within 2 weeks, Within 1 month, Anytime, DK}

Figure 33 A nonlinear model of the influence of product attributes on sales: The Screening vs. Feeling dichotomy

To handle the nonlinear influence of screening variables we used a multiplicative model:

$$Purchase = \left[\underset{\text{Screening attributes}}{\prod\left(\alpha_i + (1-\alpha_i)S_i\right)}\right] \times \left[\underset{\substack{\text{Feeling attributes}\\ \text{and other variables}}}{\sum a_j F_j}\right] \qquad 0 \le \alpha_i \le 1$$

where *Purchase* is the binary variable indicating whether or not the shopper purchased a TV, S_i is a binary variable indicating whether the Screening attribute was satisfied, F_j is a Likert-scaled or continuous Feeling attribute (or other variable in the model), and a_j are the usual regression coefficients. The α_i coefficients have a special interpretation: they represent how much the shopper cares about the corresponding Screening attribute. If α_i = 1, say for Cabinet Color, it follows that this screening variable will have no effect because the corresponding multiplicative term reduces to a constant, 1. In other words, if the shopper does not find any TV whose cabinet's color is satisfactory, they'll go ahead and still consider purchasing one of the TVs. On the other hand if α_i = 0, it immediately implies that the shopper will leave the store without purchasing a TV, regardless of all other variables in the model (including Price, Promotion, etc., not just the other product attributes). The α_i variables were estimated via nonlinear regression, not obtained from survey respondents, and thus are a true reflection of the importance of each screening attribute in terms of influence the purchase outcome. The nonlinear function described above produced significant improvement in R^2 over a linear model.

The model found that the Screening attributes have a much bigger effect on sales than the Feeling attributes. "*Not considering even a single TV at the retailer*" and "*Liking for the TV*" both ended up on our Top Ten list of critical factors, but the former has a much larger effect.

To consumers, the most important Screening attributes are *TV Type*, *Size, Specific Dimensions,* and *Brand*. *Price Range* is also fairly important, but *Cabinet color* and *Other major features* play only a small role. If consumers don't find a TV that meet their requirements on any of the important Screening attributes listed above, they will buy from somebody else.

If our client improved its performance on three of these attributes, *Type, Size*, and *Price Range*, and managed to bring the failure rate on each of these attributes down to 5%, sales would rise by 20%. In the extreme case of every shopper finding at least one TV that matches their requirements on the above attributes, sales would rise by 50%.

4.4.3.2 Advertising

While flyers drive visits by increasing awareness, it is worth noting that the *content* of advertising flyers does not appear to have much effect on visits. Flyer frequency has a nonlinear effect that diminishes surprisingly rapidly. We concluded that it is a better investment to ensure that people who don't get flyers today receive them.

4.4.3.3 Store location

Convenience of getting to a store played only a minor role in terms of driving visits, but proximity to other stores visited by the shopper had a much greater effect. For example visits to Sears for TVs increased if the store was located near a Best Buy. The triggering store did not have to sell electronics; proximity to a frequently visited grocery store had a similar effect.

4.4.3.4 Prior purchasing

Previous purchasing activity (of any item, not just TVs) at a retailer has an effect beyond driving visits: it increases the likelihood of a visitor purchasing from that retailer. Often loosely conceived as `loyalty', more than one mechanism may actually be involved, e.g., triggering (reminders, awareness), purchase confidence, etc.

4.4.3.5 Promotions

A shopper who finds their most-liked TV is on sale is twice as likely to buy as one who doesn't. We found that our client could achieve a 7% increase in sales by raising the odds of perceiving `TV on sale' by 10%

4.4.3.6 Salespeople

Salespeople play a critical role in driving TV sales: two of our top ten factors are *Talking with shoppers*, and *Explaining TV terminology*. If no salesperson talks to a customer, the likelihood of purchase goes down significantly.

4.4.3.7 Prior expectations; Store signage

Insights come from learning what is *not* important as well as learning the critical factors. Expectations (prior to visit) of finding a large product *Selection*, *High quality TVs*, or *Price Matching* at Sears had small or no effect on driving Visits.

Variables measuring store organization and signage (e.g., the *grouping* of TVs, *signs* that marked the different types of TVs, the presence or *absence of tags*, and the *information* on the tags) had small or no correlation with sales.

4.4.3.8 Online shopping

TVs were overwhelmingly bought in `bricks & mortar’ stores. While 26% of our sample went online to look for TVs and 60% of them used portals such as Yahoo and Google, less than 3% of our entire sample bought online. In general, variables measuring the online experience had little or no correlation with sales.

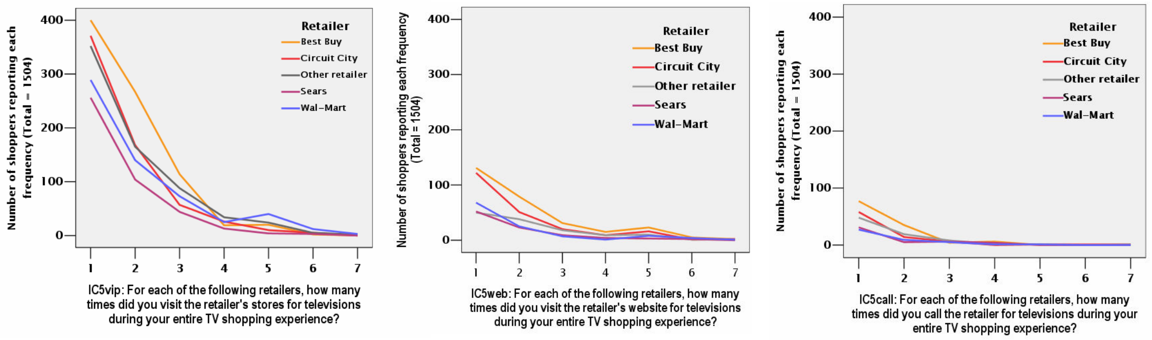


Figure 34 Frequency of Internet use

4.4.4 Model stability and generalization to other domains

A commonly asked question is how much can our findings be generalized to other domains, e.g., sales of electronics products other than TVs, non-electronics products such as appliances and groceries, and so on? A related question is how stable is the model over time, even assuming we stay within the domain of TV sales? As retailers are quick to point out, there is enormous price compression within the electronics industry, and the price of a given TV can halve within a year. What are the limits of inferences from the model?

The answer in general is “it varies for different parts of the model”. Some parts of the model are extremely robust in terms of stability over time and generalizability of results; the findings would arguably be the same a decade into the future. Other parts change rapidly and would need to be recalibrated, perhaps every year. There are two general principles which can be used to understand this.

The first principle derives from the observation that people change extremely slowly; some characteristics of human nature change only over evolutionary timescales. A person who is a bargain hunter today will probably be a bargain hunter all their life and will display similar sensitivity to price and promotion even if the study were repeated decades into the future. Therefore aspects of the model that relate to the human element, e.g., the relative importance placed by the shopper on the various product attributes, the influence of salespeople on the shopper, etc., should be quite stable.

On the other hand, variables relating solely to the product, the technology, or the retailer, such the typical price of a TV for a given type and size, can change very rapidly. Therefore the mean values and distributions of these variables need to be updated frequently.

The second principle used to understand model stability is that even when variables change rapidly over time, the *influence* (i.e., the causal effect or `importance') of the variables on another variable can be very stable. This is best understood via the illustration in Figure 35. Although the values of variable *X* can be quite unstable, changing rapidly from time t_1 to time t_2, the slope of the line is constant. Thus the amount of change in *Y* for a unit change in *X* is still the same at both points in time. Since the effect of a variable is represented by the slope, we would obtain the same `model' regardless of which point in time we conducted the study. If *X* represented price, and *Y* represented purchase, we would say that the influence of price on purchase is unchanged, even though the actual prices have changed considerably. Put another way, the *operating region* of the variables has changed; the average value of *X* has decreased from time t_1 to time t_2 (and *Y* has changed along with it). Clearly, a change in operating region does not imply a change in influence. So the model's inferences about the critical factors would remain unchanged. In short, it is extremely important not to confuse rapidly changing variables (`instability in operating regions') with changes in the model (`instability in path coefficients and causal effects). Coefficients and causal effects can also change, but this constitutes deep and fundamental change in the *mechanisms* at work in the domain, and is likely to be much slower. We believe that the findings from our model are much more robust than one might initially guess by looking at the rapid changes in the electronics industry.

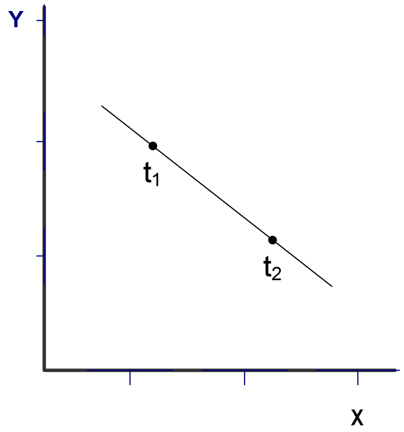


Figure 35 Stability of effects even when a variable's distribution is unstable

Generalization to other product categories is harder to assess, and is related to the question of customer segmentation. It is obvious by inspection that the purchase process is different for low-price black-box (poorly differentiated commodity-type) products such as DVD players; while TVs are viewed as furniture and looks are important, DVD players are routinely purchased sight-unseen online. Similarly there are customer segments served primarily by specialist retailers such as Ultimate Electronics and Tweeters, who purchase high-end entertainment systems costing many thousands of dollars to furnish media rooms; again the purchase process is quite different for these customer segments than the major of consumers who are served by mass-market retailers. Heterogeneity in consumer purchasing processes can also arise from behavioral characteristics intrinsic to the consumer---the same shopper who makes impulse purchases of TVs and other small purchases will also be quite impulsive in purchasing a car or a house. Apart from product and customer characteristics, the retail environment also may have characteristics that induce heterogeneity in the model, e.g., self-service environments without salespeople, such as the Internet.

All of these are general indicators of the limits of our model. To make a more specific assessment, one has to examine the model variable-by-variable and judge how much its influence on its causal descendants may be affected. One may find grounds to believe that a particular part of the model is generalizable even in the face of many forms of heterogeneity. For example, consider our finding that store layout does not have much influence on purchase because salespeople play a compensating role (Section 4.4.1). There does not appear to be anything specific to TVs in this finding; it is probably true for most other products as well. The finding can probably be generalized to domains well removed from TVs or even mass merchandize, but obviously not to a domain where there aren’t any salespeople. This probably reflects a more general principle: most of the findings are likely to be quite robust as long as any causally linked variables stay roughly within the current range. A gross movement out of range (removing all the salespeople in the store, reducing the price to nearly zero, etc.) can invalidate parts of the model by introducing severe nonlinearities.

5 Conclusion

5.1 Contributions to theory and practice

Our Consumer Purchasing Model is the most detailed and comprehensive model of factors affecting sales, that we are aware of in the marketing literature. Although it was developed specifically to explain and influence the sales of high-end televisions, much of the model is general enough to have wider applicability to other product categories and other consumer choice situations. For example, a primary finding such as the ineffectiveness of store layout in driving sales can probably be safely generalized to other high-involvement product categories and other retail environments, using the same causal mechanism as an invariant explanation, viz., that salespeople compensate for any difficulty in finding products, thereby reducing the importance of the store’s layout. The limits of such generalization are also obvious: in environments where salespeople are scarce, particularly in self-service environments, the compensating mechanism does not exist, and therefore a correlation between store layout and sales could possibly reflect a causal link. Thus researchers attempting to figure out the factors driving sales in any retail environment would benefit from testing their hypotheses against our consumer purchasing model to (a) ensure that variables that we have found to be of significant causal effect are not omitted from their study; and in particular pick up hypotheses of the causal mechanisms that may be at play; (b) better manage the scale of their study by omitting numerous other variables (e.g., store crowding) that, although commonly studied in the marketing literature, turned out to have little effect in our study; (c) guard against interpreting spurious correlations as causal effects from incorrect omission of explanatory variables (as might occur, e.g., when designing a study involving store layout variables that omits or fails to control for salespeople's behavior). The latter goal is particularly aided by the comprehensive coverage of our model, which studied potentially confounding variables from most subject areas of consumer marketing.

Another major contribution from this study arises from the modeling of quasi-deterministic relationships, the half-deterministic half-stochastic patterns that appear to be quite ubiquitous in nature. In particular, it is striking that a pattern as basic as "If you don't shop at a retailer you can't buy from them; if you do shop at them, you *may* buy from them depending on other variables" has not been explicitly handled in the literature. Instead of trying to avoid these phenomena as analytical complications arising from `missing values' in the dataset, the mathematical formulation advanced in this paper presents the researcher with a fundamental modeling construct and analytical tool that can be exploited to capture naturally occurring causal mechanisms (and in particular to capture consumer behaviors) with a significantly higher level of fidelity to the observed phenomena. Although these patterns are nonlinear in nature, the transformations that we described in this paper permit the use of mostly linear analytical tools, and given careful interpretation of the results, does not add an undue computational burden. Thus researchers interested in building statistical models of uncertain empirical phenomena (in particular, of discrete-choice decision making) who find themselves encountering deterministic and logical patterns that cause difficulties in their analyses, can now utilize one more tool along with probit regressions and other GLM tools, to expand the range of patterns that they can capture.

A third major contribution arises from the causal modeling techniques developed and tested by the study. Researchers interested in building explanatory models of any phenomenon (not just sales, but e.g., the causes of customer churn, the drivers of employee productivity, or the performance of a business), can use the quantitative methodology we described to discover causal structure in data, inject structural assumptions in a disciplined manner, build the most trustworthy model possible given the background knowledge available, and obtain a good understanding of which parts of the model are robust and which are too sensitive to the assumptions. While there are several excellent treatises available on the theory of causal modeling, getting the theory to work in practice on the noisy datasets that characterize most real-world applications, particularly those involving human behavior, continues to be a challenge. The statistical tests and algorithms described in this paper further help the practitioner in applying the right techniques and developing reliable analysis and modeling tools.

The fourth major contribution arises from how we integrated qualitative and quantitative methodologies, in particular how we fused disciplines that are normally as unconnected as ethnography and statistics. Indeed, we suggest that this approach may provide a fresh way of looking at these disciplines, and a framework for scientifically integrating them in pursuit of greater insight into consumer behavior. Key elements of this framework are the addition of structure and rigor to specific ethnographic observation methods (which ensures that variable discovery achieves a greater degree of completeness, and reduces observer-induced bias), usage of psychological metamodels to further guide variable discovery, analysis techniques used to transform raw ethnographic data into formats that support the discovery and elucidation of causal mechanisms and definition of model elements, turning these intermediate models into the basis for instrument design and quantitative data acquisition, and using the discovered causal structure in the form of supplemental assumptions to assist statistical structure discovery and model fitting. Together, these multidisciplinary techniques form a 1-2-3-4 chain of components that fit into each other like the pieces of a jigsaw puzzle. While a researcher can sometimes omit components or take shortcuts (e.g., substituting real-time ethnographic data collection with retrospective interviews), these steps provide a general framework for thinking about research methodology in a richer and deeper way than the basic qualitative/quantitative dichotomy, with a strong focus on producing trustworthy causal models in place of the more typical correlational or `indicative' results produced by non-experimental studies. Furthermore, understanding the strengths of this framework, which overcomes many of the weaknesses of traditional observational studies, empowers the researcher to design and rely more on such studies in place of experimental designs which are usually more difficult to set up, and support a much narrower range of inferences.

The fifth major contribution, and perhaps the most important one from a business perspective, is that our framework enables top managers to make the right strategic investment decisions. The underlying presence of a robust causal model takes a lot of the uncertainty out of strategy design, since it enables reliable projections of the results (or lack thereof) that would ensue from a business intervention. We illustrated this earlier with the example of the major investments in upgrading store layout that retailers currently engage in, contrasted with the finding from our causal model that any effect on sales is likely to be slim. Instead, a wiser investment would focus on the sales force, which turned out to be one of the "top ten" critical success factors. Thus, our framework enables formal "what-if" scenario analysis as the final step in applying the insight gained via the model. When a proposed business intervention happens to be directly captured by one of the variables in the model (either through foresight in study design, or by accident), the model directly computes the magnitude of the return on investment. When the proposed intervention is a creative scenario that has not been directly modeled, the methodology elicits the causal mechanisms by which the intervention is hypothesized to achieve its objectives; invariably the proposed mechanisms fully or partially intersect with those that have already been captured in the model. Thus it is possible to follow the cause-and-effect chains displayed in the model and derive an estimate of the expected return on intervention. It is also possible to work these scenarios in reverse: given a desired improvement in sales, what size of intervention will be needed to achieve the target, and given an estimated cost of the intervention, will there be a positive return on investment? Unlike the traditional `predictive' models common in marketing, the usage of causal modeling techniques thus provides a high degree of confidence in making investment decisions, and enables CEOs and Chief Marketing Officers to focus their attention on the right strategic initiatives.

5.2 Directions for further research

Some of the limitations that we outlined in Section 4.4.4 call for ways to improve our methodology, e.g., to enhance our causal modeling techniques in a way that automatically segments the model into multiple versions to improve how we deal with heterogeneity. A practical concern in building models of human behavior is the large degree of variability that we observe between individuals. For example, some shoppers are very price conscious and sensitive to promotions. Others are much more driven by recommendations from the salespeople they encounter. This sort of variability makes it harder to build a single model that adequately describes the population at large. The problem of differences between individuals is aggravated when distinctly different environments are involved: how people shop online differs significantly from how they shop in a physical store. Techniques from hierarchical Bayes and latent class modeling (e.g., [Ansari et al. 2000]) come to mind as obvious candidates for integration.

Causal modeling also engenders a powerful means of segmentation that resolves one of thorniest problems in marketing segmentation methods: it is hard to construct segments for which a sensible business action is known [Clancy and Shulman 1995]. Since actions (interventions) are the bread and butter of causal models, it is easy to add on segmentation algorithms that generate business actions as an intrinsic part of the segmentation process, rather than an afterthought. The basic idea is to look for interactions between every pair of variables in the causal model. Given two variables X_1 and X_2, if the purchase outcome Y is computed not just via the main effects (weight sum) of X_1 and X_2, but also has an interaction term, then the two variables become the basis for segmentation. Note that these variables do not have to be customer characteristics; they can be any two variables in the model. For example, if *Helpfulness of salespeople* has greater influence on *Purchase* in *Location* Philadelphia as opposed to New York,[54] it may not be necessary to train the salespeople in all locations in improve purchases. Store locations can be segmented into Philadelphia versus New York (or other appropriate regions) and training can be focused on regions where it yields greater results in terms of sales. Thus the segmentation here is that of stores + salespeople, not customers.[55] The business action is obvious and built-in: improving training in Philadelphia. Thus segmentation schemes built upon causal models can provide an extremely powerful generalization with respect to traditional segmentation methods. We have proposed one such scheme in [Noronha and Kramer 2004], but these methods need to be developed and test on real data.

Another area in which significant work is needed is software tools for causal modeling. While our Causal Modeling Workbench improves in some ways upon existing tools such as Tetrad, our system was designed for effectiveness (correctness of the model) at the expense of speed. There are many obvious ways in which our algorithms' execution speed can be improved. However the limiting factor appears to be the injection of

[54] It is commonly believed that New Yorkers have a more brusque style compared to polite Philadelphians, and customer satisfaction surveys show distinct skews in the distribution of responses.

[55] Of course, that does indirectly imply a customer segmentation, namely customers served within a given region, but that incidental; it is sufficient to think of this more directly as regional segmentation.

substantive assertions. Since the number of links in a causal model increases quadratically with the number of variables, and since a good fraction of the links cannot be given a direction by the causality detection algorithms, there is a significant manual effort required to help the algorithms along. There are many ways in which this can be improved, e.g., by developing algorithms to predict which assertions would trigger the largest number of causal inferences, and then have the user make assertions in this order so that the user's effort is minimized. Also good interactive design can make fast work of something that would otherwise be a chore, e.g., by presenting a series of assertions in a highly organized format that requires just a single click for acceptance. Our workbench is currently too basic in terms of user interface and algorithmic efficiency to handle very large models. While it has served very well for our model (which is extraordinarily large for a marketing model), we can easily conceive of much larger datasets from other domains that would need the increased efficiency.

6 Acknowledgements

We gratefully the significant contributions of many people to this study, especially Yasuo Amemiya, Steve Ballou, Mareike Barmeyer, Ken Bollen, Carolyn Brodie, Tom Cofino, Susan Confort, Myron Flickner, Mary Jo Frederich, Terese Friel, Megan Gordon, Todd Gurley, Stephanie Hardt, Mackenzie Harris, Jim Hilt, Sandra Kogan, Alex Liu, Paul Maglio, Jorge Sanz, Byron Thayer, John Vergo; Gary Van Prooyen, Liz Siegel, and the Sears team; Nina Nichols and the Resolution Research team; and Steve Barnard and the Link Group team.

7 Bibliography

1. Ralph Adoplhs, "Neural systems for recognizing emotion", *Neurobiology,* **12**:169–177, 2002.

2. I. Ajzen, "The theory of planned behavior," *Organizational Behavior and Human Decision Processes*, 50, 179-211, 1991

3. I. Ajzen, "Nature and operation of attitudes," *Annual Review of Psychology*, 52:27-58, 2001.

4. I. Ajzen, and M. Fishbein, "Attitudes and the attitude-behavior relation: Reasoned and automatic processes" In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology* (pp. 1-33). John Wiley & Sons., 2000.

5. C. T. Allen, K. Machleit and S. S. Kline, "A comparison of attitudes and emotions as predictors of behavior at diverse levels of behavioral experience," *Journal of Consumer Research,* Vol. 10, pp. 493-504, 1992.

6. Paul D. Allison, *Missing Data*, Sage Publications, Thousand Oaks, CA, 2001.

7. D. Andrews, B. Nonnecke, J. Preece, "Electronic survey methodology: A case study in reaching hard to involve Internet Users," *International Journal of Human-Computer Interaction*. 16, 2, 185-210, 2003.

8. Asim Ansari, Kamel Jedidi and Sharan Jagpal, “A hierarchical Bayesian methodology for treating heterogeneity in structural equation models,” *Marketing Science*, Vol. 19, No. 4, pp. 329-347, Fall 2000.

9. Christopher J. Armitage and Mark Connor, “Efficacy of the Theory of Planned Behavior: a meta-analytic review,” *British Journal of Social Psychology*, Vol. 40 (4), pp. 471-499, 2001

10. Eric J. Arnould and Melanie Wallendorf, “Market-oriented ethnography: Interpretation building and marketing strategy formulation,” *Journal of Marketing Research*, Vol 31, pp 484-504, Nov. 1994.

11. Stephen E. Ballou, Personal communication, 2004.

12. Steve Ballou, Julian Chu and Gina Paglucia Morrison, “Deeper customer insight: Understanding today's complex shoppers”, *IBM Institute for Business Value Study*, 2005.

13. Rajeev Batra and Olli T. Ahtola, “Measuring the hedonic and utilitarian sources of consumer attitudes,” *Marketing Letters*¸2:2, 159-170, 1990.

14. Michael P. Battaglia, Kate Ballard-LeFauve, Linda Piccinino, Mary Cay Murray, Robert A. Wright, Meena Khare, “Reducing attrition in a random digit dialing based provider recorder check study”, *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9, 2001

15. W. O. Bearden and R. G. Netemeyer, *Handbook of Marking Scales*, 2nd Edition, Sage Publications, Thousand Oaks, 1999.

16. Antoine Bechara, Hanna Damasio, D. Tranel and A. R. Damasio: “Deciding advantageously before knowing the advantageous strategy”, *Science*. 1997 Feb 28;275 (5304):1293-5.

17. Antoine Bechara, Hanna Damasio and Antonio Damasio, “Emotion, decision making and the orbitofrontal cortex”, *Cerebral Cortex.* 2000 Mar;10(3):295-307.

18. Genevieve Bell and R. Teague, “Getting Out of the Box: Ethnography Meets Life, Applying anthropological techniques to experience research”, Tutorial Notes, *UPA Conference 2001*, Lake Las Vegas, Nevada, 2001

19. M. Ben-Akiva, D. McFadden, et al., “Hybrid Choice Models: Progress and Challenges,” *Marketing Letters*, 13:3, pp. 163-175, 2002.

20. Robert P. Berrens et al., “The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples,” Political Analysis, Volume 11, Number 1, pp. 1-22, 2003.

21. Herman J. Bierens, “Maximum likelihood estimation of Heckman’s sample selection model”, [WWW document], URL http://econ.la.psu.edu/~hbierens/EasyRegTours/HECKMAN.PDF, July 2002, Retrieved Nov 1 2006.

22. Kenneth A. Bollen, “Latent variables in psychology and the social sciences,” *Annual Review of Psychology*, Vol. 53, pp. 605-34, 2002.

23. Kenneth A. Bollen, *Structural Equations with Latent Variables*, John Wiley and Sons, New York, 1989.

24. Grady Booch et al., *The Unified Modeling Language User Guide*, Addison-Wesley, Reading, Massachusetts, 1999.

25. Richard E. Boyatzis, *Transforming Qualitative Information: Thematic Analysis and Code Development,* Sage Publications, Thousand Oaks, 1998.

26. J. Michael Brick, Joseph Waksberg, "Bias in List-Assisted Telephone Samples" *American Association of Public Opinion Research,* May 14, 1994

27. Ray Burke, "Creating the ideal shopping experience," Indiana University---KPMG Study, 2000.

28. Patrick Butler and Joe Peppard, "Consumer purchasing on the Internet: Processes and prospects," *European Management Journal*, Vol. 16, No. 5, pp. 600-610, 1998.

29. Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*¸ Houghton Mifflin Company, Boston, 1963.

30. John P. Chin, Virginia A Diehl and Kent L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," *Computer Human Interaction 1988,* pp. 213-218, 1988.

31. Kevin J. Clancy, Robert S. Shulman, *Marketing Myths That Are Killing Business: The Cure for Death Wish Marketing*, McGraw-Hill, 1995.

32. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Lawrence Erlbaum, New Jersey, 1988.

33. J. B. Cohen and C. S. Areni, "Affect and consumer behavior", Thomas S. Roberston and Harold H. Kassarjian, *Handbook of Consumer Behavior,* Prentice-Hall, Englewood Cliffs, NJ, 1991.

34. A. Crabtree, J. O'Brien, D. Nichols, M. Rouncefield, and M. Twidale, "Ethnomethodologically informed ethnography and information systems design", *Journal of the American Society for Information Science*, vol. 51 (7), pp. 666-682, 2000.

35. Antoine R Damasio, *Descartes' Error : Emotion, Reason and the Human Brain,* Avon Books, 1994.

36. Richard B Darlington, "Factor Analysis", [WWW document] URL http://comp9.psych.cornell.edu/Darlington/factor.htm, Retrieved Nov 1 2006.

37. Don A. Dillman, *Mail and Internet Surveys*: *The Tailored Design Method*, John Wiley and Sons, New York, 2000.

38. R. J. Donovan, and John R. Rossiter, "Store Atmosphere: An Environmental Psychology Approach," Journal of Retailing, 58 (Spring), 34-57, 1982.

39. N. Duan, W. G. Manning, C. N. Morris, and J. P. Newhouse (1983) A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics*, 1, 115-126.

40. A. H. Eagly and S Chaiken, *The Psychology of Attitudes,* Harcourt Brace Jovanovich College Publishers, Fort Worth, TX, 1993

41. Jeffrey R Edwards and Richard P Bagozzi, "On the nature and direction of relationships between constructs and measures", *Psychological Methods*, Vol 5, No. 2, p.155-174, 2000.

42. Robert M. Emerson, *Writing Ethnographic Fieldnotes*, The University of Chicago Press, Chicago, 1995.

43. Sunil Erevelles, "The role of affect in marketing," *Journal of Business Research*, 42, 199-215, 1998.

44. K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data,* MIT Press, Cambridge, MA., 1993.

45. David Freedman, "From association to causation via regression", *Notre Dame Conference on Causality in Crisis,* Oct 15-17, 1993.

46. David Freedman, "Are there algorithms that discover causal structure", Technical Report 514, Department of Statistics, University of California, 30 June 1998.

47. H. H. Friedman and Taiwo Amoo, "Rating the rating scales," *Journal of Marketing Management*, Vol. 9, No. 3, pp. 114-123, Winter 1999.

48. Robert D. Froman, "Elements to consider in planning the use of factor analysis," *Southern Online Journal of Nursing Research*, Vol. 2, No. 5, 2001. [WWW document] URL http://www.snrs.org/publications/SOJNR_articles/iss05vol02.pdf Retrieved Nov 1 2006.

49. S. F. Gardial, et al. "Comparing consumers' recall of prepurchase and postpurchase product evaluation experiences," *Journal of Consumer Research,* Vol 20, March 1994.

50. David Garson, "Structural equation modeling," PA765 Course notes, NCSU, [WWW document] URL http://www2.chass.ncsu.edu/garson/pa765/structur.htm Retrieved Nov 1 2006.

51. Hye-Young Kim and Youn-Kyung Kim, "Escapism, Consumer Lock-in, Attitude, and Purchase: An Illustration from an Online Shopping Context," *Journal of Shopping Center Research*, Volume 12, Issue 2, 2005.

52. Lee H. Giesbrecht, Dale W. Kulp and Amy W. Starer, "Estimating Coverage Bias in RDD Samples with Current Population Survey (CPS) Data", *American Statistical Association (ASA) Conference*, August 1996.

53. Joan L. Giese and Joseph A. Cote, "Defining consumer satisfaction," *Academy of Marketing Science Review*, Vol 2000, No. 1, 2000.

54. Ted Goertzel, "Myths of Murder and Multiple Regression: Econometric Modeling as Junk Science," *The Skeptical Inquirer*, Volume 26, No 1, January/February 2002, pp. 19-23.

55. Abbie Griffin and John R. Hauser, "The voice of the customer," *Marketing Science*, Vol. 12 No. 1, pp. 1-26, Winter 1993.

56. Sunil Gupta and Valarie Zeithaml, “Customer metrics: the past, the present and the future in academia and practice,” *Marketing Science Institute Report* 05-200, 2005.

57. John R. Hauser and Vithala Rao "Conjoint Analysis, Related Modeling, and Applications," *Advances in Marketing Research: Progress and Prospects*, Jerry Wind, Ed., 2003.

58. W. G. Hopkins, *A New View of Statistics*, http://newstatsi.org, 2000. Retrieved June 2006.

59. S. R. Hutchinson, “The stability of post hoc model modifications in confirmatory factor analysis models,” *The Journal of Experimental Education*, vol. 66, no4, pp. 361-380 1998.

60. Tsuyoshi Idé, “Generalizing covariance using group theory,” *ICDM* 2005.

61. Magid Igbaria and Saroj Parasuraman, “Attitudes toward microcomputers: development and construct validation of a measure,” *International Journal of Man-Machine Studies,*Vol 35, pp. 553-573, 1991.

62. Janet Ilieva, Steve Baron and Nigel M Healey, “On-line Surveys in International Marketing Research: Pros and Cons,” Manchester Metropolitan University Business School Working Paper Series, WP01/10, July 2001.

63. Blake Ives, M. H. Olson and J. J. Baroudi, “The measurement of user information satisfaction,” *Communications of the ACM*, Vol. 26, No. 10, pp. 785-93, 1983.

64. Carroll E. Izard, *Human Emotions*, New York: Plenum Press, 1977.

65. J. Jacoby, G. V. Johar and M. Morrin, “Consumer behavior: a quadrennium,” *Annual Review of Psychology*, 49:319-44, 1998.

66. Cheryl B. Jarvis, Scott B. Mackenzie and Philip P. Podsakoff, “A critical review of construct indicators and measurement model misspecification in marketing and consumer research,” *Journal of Consumer Research*, Vol. 30, No. 2, Sep 2003, pp. 199-218.

67. Michael P Jones, “Indicator and stratification methods for missing explanatory variables in multiple linear regression,” *Journal of the American Statistical Association*, Vol. 91, No. 433, March 1996, pp. 222-230.

68. Daniel Kahneman, “Evaluation by moments: past and future,” D. Kahneman and A. Tversky, *Choices, Values and Frames*, Cambridge University Press, New York, 2000.

69. Jack Katz, “From how to why: On luminous description and causal inference in ethnography (Part 1)”, *Ethnography*, Vol 2(4): 443-473, 2001.

70. Jack Katz, “From how to why: On luminous description and causal inference in ethnography (Part 2)”, *Ethnography*, Vol 3(1): 63-90, 2002.

71. Ralph L. Keeney, “The value of Internet commerce to the customer,” *Management Science*, Vol. 45, No. 4, pp. 533-542, April 1999.

72. Jurek Kirakowski, “The Use of Questionnaire Methods for Usability Assessment” 2001, [WWW document] Retrieved 02 Nov 2001, URL http://www.ucc.ie/hfrg/questionnaires/sumi/sumipapp.html

73. Jed Kolko, Chris Charron, Sheila Baxter, “How to sell consumer electronics”, *Forrester Technographics Report*, September 2003.

74. Philip D. Kotler, *Marketing Management*, Prentice Hall, Upper Saddle River, New Jersey, 2000.

75. Joseph D. Kramer and Sunil J. Noronha, “Understanding the Customer Experience: Critical Success Factors Linking Customer Experience to Business Drivers”, Sears FOAK Phase 1 Report, IBM, July 2003.

76. Joseph D. Kramer and Sunil J. Noronha, “Understanding the Customer Experience: Critical Success Factors Linking Customer Experience to Business Drivers”, Sears FOAK Final Report, IBM, December 2005.

77. Jon A. Krosnik, “Survey research,” *Annual Review of Psychology*, Vol. 50, pp.537-67, 1999.

78. Jon A. Krosnick and LinChiat Chang, “A Comparison of the Random Digit Dialing Telephone Survey Methodology with Internet Survey Methodology as Implemented by Knowledge Networks and Harris Interactive”, April 2001, [WWW document] URL http://www.knowledgenetworks.com/ganp/docs/OSUpaper.pdf Retrieved Nov 1 2006.

79. Lakoff, G. and Johnson, M., *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought,* New York: Basic Books, 1999.

80. R. C. MacCallum et al., “Model modifications in covariance structure analysis: the problem of capitalization on chance,” *Psychol Bull*. May;111(3):490-504, 1992.

81. Mary J. LaLaomia and Joseph B. Sidowski, “Measurements of computer satisfaction, literacy, and aptitudes: a review,” *International Journal of Human-Computer Interaction*, Vol 2, No. 3, pp. 231-253, 1990.

82. Gary L. Lilien, Philip Kotler, and K. Sridhar Moorthy, *Marketing Models*¸ Prentice-Hall, Upper Saddle River, New Jersey, 1992.

83. M Limayem, M. Khalifa and A. Frini, “What makes consumers buy from the Internet? A longitudinal study of online shopping,” *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 30, No. 4, July 2000.

84. Gerald Lohse, “Usability and profits in the digital economy,” in McDonald, S., Waern, Y. and Cockton, G. (eds.). *People and Computers XIV - usability or else. Proceedings of HCI 2000.* London : Springer - Verlag. pp 3 – 16, 2001.

85. H. Mano, and Richard L. Oliver "Assessing the dimensionality and structure of the consumption experience: Evaluation, feeling and satisfaction", *Journal of Consumer Research*, 20 (December), 451-66, 1993.

86. Albert Mehrabian, and James Russell, *An Approach to Environmental Psychology*. Cambridge, Mass.: MIT Press, 1974.

87. Mathew B. Miles and A. Michael Huberman, *Qualitative Data Analysis*, Sage Publications, Thousand Oaks, CA, Second Edition, 1994.

88. Jeff Miller and Dan Coates, “Online survey length: can research findings be impacted,” *ARF Annual Convention and Infoplex*, New York, April 9 2003.

89. Thomas W. Miller and Peter R. Dickson, “On-line market research,” *International Journal of Electronic Commerce*, Vol. 5, No. 3, pp. 139-167, Spring 2001.

90. John C. Mowen and Michael Minor, *Consumer Behavior*, Prentice-Hall, Upper Saddle River, NJ, 1998.

91. B. O. Muthén, *Mplus Technical Appendices*, Los Angeles, CA: Muthén & Muthén, 1998-2004.

92. Gad Nathan, “Telesurvey methodologies for household surveys---a review and some thoughts for the future,” *Survey Methodology* 27, pp.7-31, 2000.

93. Sunil J. Noronha and Joseph Kramer, “System and Process for Discovering Factors that Influence Decisions and other Behavioral Outcomes,” Patent disclosure, IBM, March 2004.

94. Sunil J. Noronha and Joseph Kramer, “System and Method for Building Cognitive Category Based Mental Models,” Patent disclosure, IBM, January 2006.

95. T. P. Novak, D.L. Hoffman, and Y.F. Yung, "Measuring the Customer Experience in Online Environments: A Structural Modeling Approach," *Marketing Science*, Winter, 19(1), 22-44, 2000.

96. T. P. Novak, D. L. Hoffman and A. Schlosser, “Consumer control in online environments,” 2000, [WWW document] URL http://advertising.utexas.edu/vcbg/home/Hoffman00.pdf Retrieved Winter 2001.

97. R. L. Oliver, “Cognitive, affective, and attribute bases of the satisfaction response,” *Journal of Consumer Research*, Vol 20., Dec 1993

98. T. M. Ostrom, “Interdependence of attitude theory and measurement” In A. R. Pratkanis, S., J. Breckler & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 11-36). Hillsdale, NJ: Erlbaum, 1989.

99. Rodrigo A. Padilla “Literature review on: Consumer satisfaction in modern marketing”, 1996 [WWW document] URL http://pages.infinit.net/rodrigo/satisfaction.html Retrieved Jan 28 2002.

100. Judea Pearl, *Causality*, Cambridge University Press, Cambridge, UK, 2000.

101. Richard E. Petty et al., “Attitudes and attitude change,” *Annual Review of Psychology*, 48:609-47, 1997.

102. A. R. Pratkanis, S. J. Breckler and A. G. Greenwald, (ed.) *Attitude Structure and Function*, Hillsdale, N. J., 1989.

103. S. Rajgopal, M Ventatachalam and S Kotha, “Does the quality of online customer experience create a sustainable competitive advantage for e-commerce firms?" *Social Science Research Network Working Paper*, December 20, 2000. [WWW document] URL http://ssrn.com/abstract=242774

104. G. Rees, G. Kreiman, and C. Koch, “Neural correlates of consciousness in humans”, *Nat Rev Neurosci*. 2002 Apr;3(4):261-70.

105. Retail Forward, Consumer Electronics Shopper Update, March 2004.

106. Thomas J. Reynolds and David B. Whitlark, “Applying laddering data to communications strategy and advertising practice,” *Journal of Advertising Research*, pp. 9-17, July/August 1995.

107. Marsha L. Richins, “Measuring emotions in the consumption experience,” *Journal of Consumer Research*, Vol 24, Sep 1997.

108. James K. Rilling, David A. Gutman, Thorsten R. Zeh, Giuseppe Pagnoni, Gregory S. Berns, and Clinton D. Kilts, “A Neural Basis for Social Cooperation”, *Neuron*, Vol. 35, 395–405, July 18, 2002

109. A. Rushinek and S. F. Rushinek, “What makes users happy?” *Communications of the ACM*, Vol. 29, No. 7, 594-8, 1986.

110. Joseph L. Schafer and John W. Graham, Missing Data: Our View of the State of the Art, *Psychological Methods*, Vol. 7, No. 2, 147–177, 2002.

111. Joseph L. Schafer Maren K. Olsen, “Modeling and imputation of semicontinuous survey variables,” Proceedings *of Federal Committee on Statistical Methodology* September 1999.

112. Jeffrey D. Schall, “The neural basis of deciding and acting”, *Nature Reviews Neuroscience*, Vol 2., p.33-42, Jan 2001.

113. Helen B. Schwartzman, *Ethnography in organizations,* Sage Publications, Newbury Park, Calif. 1993.

114. Norbert Schwarz, “Self-reports: How the questions shape the answers,” *American Psychologist,* Vol. 54, No. 2, pp. 93-105, February 1999.

115. Norbert Schwarz and Gerald L. Clore, “Feelings and phenomenal experiences,” in E. Tory Higgins and Arie W Kruglanski (ed.) *Social Psychology: Handbook of Basic Principles*, The Guilford Press, New York 1996.

116. Michael Scriven, “The logic of evaluation,” 2002 [WWW document] URL http://eval.cgu.edu/lectures/intro/notewk1b.htm#logic of evaluation, Retrieved Dec 2002.

117. Venkatesh Shankar, Amy K. Smith, and Arvind Rangaswamy, “Customer satisfaction and loyalty in online and offline environments,” International Journal of Research in Marketing, Vol. 20, 2003.

118. J. A. Simpson and E. S. C. Weiner, *The Oxford English Dictionary*, Clarendon Press, Oxford, 1998.

119. Peter Spirtes, Thomas Richardson, Chris Meek, Richard Scheines, Clark Glymour, “Using path diagrams as a structural equation modeling tool” *Sociological Methods and Research*, 27(2), November 1998.

120. Peter Spirtes, Clark Glymour and Richard Scheines, *Causation, Prediction, and Search*, The MIT Press, Cambridge, Massachusetts, 2000.

121. Peter Spirtes et al., *Tetrad III* and *Tetrad IV*, 2004, [Software] URL http://www.phil.cmu.edu/projects/tetrad/index.html, Retrieved Nov 1 2006.

122. James P. Spradley, *The Ethnographic Interview*, Harcourt Brace Jovanovich College Publishers, Fort Worth, 1979.

123. James P. Spradley, *Participant Observation*, Harcourt Brace Jovanovich College Publishers, Fort Worth, 1980.

124. Daniel Steel, "Genetic Redundancy and the Faithfulness Condition", presented at the meeting of the British Society for Philosophy of Science, University of Kent at Canterbury, July 8, 2004.

125. Willam R. Swinyard, "The effects of mood, involvement and quality of store experience on shopping intentions," *Journal of Consumer Research*¸ Vol 20, September 1993.

126. Abraham Tesser and Leonard Martin, "The psychology of evaluation," in E. Tory Higgins and Arie W Kruglanski (ed.) *Social Psychology: Handbook of Basic Principles*, The Guilford Press, New York 1996.

127. U.S. Department of Housing and Urban Development, "Random Digit Dialing Surveys: A Guide to Assist Larger Housing Agencies in Preparing Fair Market Rent Comments", Office of Policy Development and Research, Economic and Market Analysis Division, April 2000.

128. P. M. West, P. L. Brockett and L. L. golden, "A comparative analysis of neural networks and statistical methods for predicting consumer choice," *Marketing Science*, Vol. 16, No. 4, 1997.

129. Robert A. Westbrook, Intrapersonal affective influences on consumer satisfaction with products," *Journal of Consumer Research*, Vol 7., June 1980.

130. Wikipedia, "Grounded theory (Glaser)" [WWW document] URL http://en.wikipedia.org/wiki/Grounded_theory_%28Glaser%29, Retrieved Nov 1 2006.

131. Jochen Wirtz and Meng Chung Lee, "An Examination of the Quality and Context-Specific Applicability of Commonly Used Customer Satisfaction Measures," *Journal of Service Research*, Vol. 5, No. 4, 345-355 2003.

132. Stephen Worchel, *Social Psychology*, Wadsworth, Australia, 2000.

133. Xiao-Hua Zhou and Hua Liang, "Semi-parametric Single-index Two-Part Regression Models, UW Biostatistics Working Paper Series, Paper 235, University of Washington, 2004.

134. M. P. Zanna, and J. K. Rempel, "Attitudes: A new look at an old concept. In D. Bar-Tal, & A. W. Kruglanski (Eds.), The *Social Psychology of Knowledge*. (pp. 315-334), 1988.