# Scale Matters:
# How Craft in Conjoint Analysis Affects Price and Positioning Strategies

by

Felix Eggers

John Hauser

and

Matthew Selove

**June 2017**

Felix Eggers is an Assistant Professor of Marketing and Fellow of the SOM Research School at the Faculty of Economics and Business, University of Groningen, Nettelbosje 2, 9747 AE Groningen, The Netherlands. +31 50 363 7065, f.eggers@rug.nl.

John R. Hauser is the Kirin Professor of Marketing, MIT Sloan School of Management, Massachusetts Institute of Technology, E62-538, 77 Massachusetts Avenue, Cambridge, MA 02142, (617) 253-2929, hauser@mit.edu.

Matthew Selove is an Associate Professor of Marketing at The George L. Argyros School of Business and Economics, Chapman University, Beckman Hall 303L, 1 University Drive, Orange, CA 92866, selove@chapman.edu.

We are grateful to Greg Allenby, Tammo Bijmolt, Daria Dzyabura, Joel Huber, Tony Ke, Martin Meissner, Olivier Toubia, and Jaap Wieringa for their constructive comments.

# Scale Matters:
# How Craft in Conjoint Analysis Affects Price and Positioning Strategies

## Abstract

Managerial decisions frequently depend on choice-based conjoint (CBC) studies. It is well-accepted that the relative partworths need to be estimated accurately to be of managerial value. We illustrate that managerial recommendations from CBC also depend critically on the (observed) absolute scale of the partworths, even if the relative partworths are accurate. Greater scale means that the partworths are larger relative to random errors. Researchers face the challenge of identifying the scale that represents consumers' actual choices when making purchase decisions. In eight CBC studies, we vary the realism of stimuli, incentive alignment, training videos, and detail of instructions and thus demonstrate that the quality of execution of a CBC study ("craft"), and the type of validation task, both affect observed scale (and managerial decisions) substantially. Failure to strive for the highest possible craft and failure to evaluate scale with respect to realistic validation tasks may lead to costly suboptimal managerial pricing and positioning decisions. If patent or copyright valuations are based on low craft studies, they can be off by hundreds of millions of dollars.

Keywords: *Conjoint analysis, willingness to pay, choice modeling, surveys*

## *MOTIVATION*

Conjoint analysis has become the preferred method to measure consumer preferences for product or service attributes with over 18,000 applications per year—the vast majority of them (93%) with a form of choice-based conjoint analysis (CBC, Orme 2009, p. 127, Orme 2016). General Motors alone spends tens of millions of dollars each year on conjoint analysis studies (Urban and Hauser 2004). Given industry interest, it is not surprising that academics and practitioners publish many papers each year exploring improved estimation methods, adaptive question selection, and methodological variations (reviews in Orme 2009 and Rao 2014).

In defining the quality of a CBC study, researchers and managers typically focus on accuracy, where, by accuracy, we mean the ability of the study to identify the true relative utility weights, known as partworths. For example, if a consumer values a change from a gold- to a silver-colored smartwatch twice as much as a change from brown- to a black-leather band, then we call a CBC study accurate if the estimated relative partworths are also in a two-to-one ratio. Accuracy is clearly important managerially because accuracy helps firms decide which product attributes consumers prefer.

We argue that accuracy in the relative partworths is a necessary, but not sufficient, condition for drawing correct managerial implications from CBC studies. The size of the error term relative to the size of the partworths also plays an important role. When partworths of the modeled attributes explain more of the uncertainty in consumer choice, the absolute magnitude of the partworths increases. Following Train (2006, p. 44), we call the absolute magnitude "scale." (We define accuracy and scale formally in the next section.) For example, if a consumer is distracted when making decisions and often chooses an alternative that does not represent his or her true maximum utility, scale leans towards zero and choice predictions are less likely to differentiate among alternatives. On the other hand, if the attributes and alternatives are described realistically and the consumer is motivated to put the same effort into the choice tasks as in the marketplace, the scale of the partworths will be closer to their true values and the choice probabilities will reflect better how the consumer will react to real products. To the extent that there is inherent stochasticity in consumer decision making, probabilities are not 0 or 1 and scale is not infinite. Thus, rather than artificially inflating scale or ignoring the error term with "first-choice" simulators (Sawtooth Software 2005), researchers face the challenge of identifying the scale of the partworths that represents consumers' actual choices when making purchase decisions.

In this paper, we separate scale and accuracy and show that scale matters when estimating the market equilibrium price of products and when making strategic decisions about product attribute levels. Specifically, we explore whether the quality of execution of a CBC study, also known as "craft,"

affects observed scale and, in turn, strategic decisions. For example, in practice, text-based stimuli are common, incentive-alignment is not common, and training videos are rarely used. Practitioners justify their decisions because replacing text with animations, adding incentive alignment, or creating training videos is too costly. Software defaults are deemed sufficient (e.g., Orme 2009; Sawtooth Software 2014, 2015; Thaler and Sunstein 2009). Our experience suggests that practitioners undervalue investments in craft because CBC practice focuses primarily on accuracy. Although there has been substantial research investigating how heterogeneity in scale affects the interpretation of CBC analyses (e.g., Louviere 2001 and Salisbury and Feinberg 2010), there has been less focus on the managerial relevance of observed scale.

We undertake eight CBC studies using a professional panel to explore empirically how four key elements of craft affect observed scale, accuracy, and managerial decisions. The key elements are realism of the stimuli, incentive alignment, enhanced instructions, and training videos, varied in a half-replicate design. We find that realistic stimuli (vs. text-only stimuli) and incentive alignment affect observed scale to such a degree as to change managerial decisions. When CBC is used to evaluate patents and copyrights (e.g., Allenby, et al. 2014b), the effect of craft on scale can account for hundreds of millions of dollars in valuations. Video instructions and instructions to hold "all else equal" have less of an effect.

Perhaps less attention is focused on scale because CBC models are primarily evaluated on measures of *internal* (holdout) validity rather than *external* validation such as consumer tasks that attempt to better mimic marketplace choices. Measures based on internal holdout tasks can be misleading when assessing scale, because consumers can learn to answer choice tasks consistently, resulting in a larger observed (estimation-based) scale, whether or not the choice tasks mimic external measures. We test a scale-adjustment based on a delayed, market-like validation task and evaluate the impact on observed scale, as based on the estimated partworths, versus observed scale, as adjusted to mimic marketplace decisions. We show that managerial pricing and positioning decisions depend strongly on whether or not the observed scale is adjusted. Moreover, we show that investments in craft are able to reduce this discrepancy.

We next conceptualize and formalize scale and accuracy. We distinguish observed scale upon which the firm bases its decisions, from the true scale upon which the marketplace reacts. Subsequent sections demonstrate that scale, true and/or observed, affects predicted equilibrium prices, strategic decisions about product attribute levels, and the valuation of patents/copyrights. We then describe, analyze, and interpret the empirical studies. By demonstrating that the effect of craft on managerial

recommendations can be dramatic, we hope to spark research to explore craft in general.

## *CONCEPTUALIZATION AND FORMALIZATION OF SCALE AND ACCURACY*

### *Definition of Scale*

The standard CBC model is based on a linear-in-the-parameters random-utility interpretation. Let $\vec{a}_j$ be a vector of product attributes for product profile $j$. For example, a smartwatch profile can be either gold- or silver-colored, and can have either a black- or a brown-leather band. (Our notation can be modified easily for multi-level attributes, interactions of attributes, and non-linear functions of attributes.) Let $\vec{\beta}_i$ be the vector of weights that consumer $i$ places on the attributes, i.e., the partworths. Let $p_j$ be the price of product profile $j$, let $\eta_i$ be the weight that consumer $i$ places on price, and let $u_o$ be the relative utility of the outside option.

If $\epsilon_{ij}$ is an independent and identically distributed Gumbel error with variance $\pi^2/(6\gamma_i^2)$, then CBC models consumer $i$'s utility for the $j^{th}$ profile by Equation 1 ( $'$ indicates vector transpose).

(1) $$u_{ij} = \vec{\beta}_i{}' \vec{a}_j - \eta_i p_j + \epsilon_{ij}$$

Using this formulation, Ben-Akiva and Lerman (1985, p. 105) show that the probability, $P_{ij}$, that consumer $i$ will choose profile $j$ from a choice set containing $J$ products and an outside option, is given by a logit model:

(2) $$P_{ij} = \frac{e^{\gamma_i(\vec{\beta}_i{}' \vec{a}_j - \eta_i p_j)}}{e^{\gamma_i u_{i_o}} + \sum_{k=1}^{J} e^{\gamma_i(\vec{\beta}_i{}' \vec{a}_k - \eta_i p_k)}}$$

Train (2006, p. 44-45) calls $\gamma_i$ the scale factor suggesting that only the combined values, $\gamma_i\vec{\beta}$ and $\gamma_i\eta_i$, can be identified—we cannot simultaneously identify $\gamma_i, \eta_i$, and $\vec{\beta}_i$. See also Miller, et al. (2011). In standard logit models, including CBC, practitioners routinely impose a linear constraint for identification. One constraint sets the error variance to $\pi^2/6$ such that the empirical model estimates observed partworths: $\vec{\beta}_i^{observed} \equiv \gamma_i\vec{\beta}_i$ and $\eta_i^{observed} \equiv \gamma_i\eta_i$. Alternatively, McFadden (2014) sets $\eta_i = 1$ so that the $\vec{\beta}_i$'s are interpreted relative to price. However, such scaling is only appropriate when the relative partworths, $\vec{\beta}_i$, do not vary among experimental conditions. In order to compare the effect of craft, which may affect accuracy as well as observed scale, a more informative constraint sets the $\beta_i$'s such that the difference between the most preferred and least preferred alternative is 1. Observed scale is then the sum of importances based on the $\vec{\beta}_i^{observed}$'s. (The importance of an attribute is the largest partworth for a level of the attribute minus the smallest partworth for a level of the attribute.) With this

definition of scale, the absolute magnitudes of the observed partworths ($\vec{\beta}_i^{observ\,ed}$'s) are larger when scale is larger. When relative partworths do not vary among experimental conditions, this definition of scale is proportional to a definition based on McFadden's constraint. Another interpretation of scale is that the magnitude of observed utility (the partworths) is larger relative to the standard deviation of the error term (e.g., Arora and Huber 2001).

### *Estimation, Validation, and True Scale*

Empirically, true scale and true accuracy cannot be observed directly. Instead, researchers typically evaluate the (internal) predictive ability of a CBC model by comparing predictions to holdout choice tasks. However, evaluating observed scale and accuracy based on holdout tasks might not represent how consumers behave in the actual marketplace. For example, suppose all attributes are described by text-only in the CBC experiment. Consumers might be extremely consistent in making choices among text-only product profiles—both in the choice sets and in the holdout task. A model estimated from ten choice sets might predict holdout choices well—the observed scale based on the estimated partworths might be extremely high. In fact, internal validity and observed scale might be higher for text-only stimuli than for more realistic stimuli. However, a model based on more-realistic stimuli might represent marketplace choices better than one based on text-only stimuli. Partworths from a CBC study based on more-realistic stimuli might predict choice in a market-like validation task (separate from the CBC choice sets) better than those from a text-only CBC study, leading to higher external validity for the study based on realistic stimuli.

Let the $\gamma_i^{true}$ be the scale that accounts for how the consumer will actually react to changes in product attribute levels and price in the marketplace. Empirically, even with a full specification of the CBC model and the highest craft, there remains residual stochasticity in the consumer choice—a phenomenon that is well-documented in the consumer behavior and marketing science literatures. For simplicity of notation, let $\gamma^{true}$ refer to the set of the $\gamma_i^{true}$'s. A CBC study seeks to estimate observed scale values that are close to $\gamma^{true}$. We hypothesize that lower-craft studies, e.g., text-based stimuli rather than rich pictures and animations, lead to observed scale values, $\gamma^{lower}$, that are further from the $\gamma^{true}$ than observed scale values that are based on higher craft, $\gamma^{higher}$. Mathematically, we expect that $\left\|\gamma^{true} - \gamma^{higher}\right\| < \left\|\gamma^{true} - \gamma^{lower}\right\|$, where the norm is defined over $i$. ($\|\ldots\|$ indicates a vector-space norm.) The superscripts refer to higher and lower craft, not higher or lower values of scale. We test these hypotheses in our empirical studies.

By definition, consumer reaction in the marketplace is determined by the $\gamma^{true}$. However, the

firm's decisions are based on the firm's predictions of marketplace response; the firm will act on the observed scale reported by the CBC study, $\gamma^{observed}$. Suppose a firm observes $\gamma^{lower}$ in a lower-craft study, which may be larger or smaller than $\gamma^{true}$. If $\gamma^{lower}$ is quite different than $\gamma^{true}$, then the firm might make strategic errors when it selects its price and attribute levels. As a potential remedy to avoid such strategic errors, we might seek to adjust the observed scale to reflect the ability of the CBC model to predict a measure that mimics marketplace choices. We call this validation-adjusted scale, $\gamma^{validation}$. The better the validation measure mimics marketplace choices, the closer $\gamma^{validation}$ is to $\gamma^{true}$. Empirically, we find the effect is dramatic and justifies research to improve measures of external validity, perhaps even beyond our marketplace-mimicking measure.

### *Accuracy*

Conditioned on scale, we say the partworths are more accurate the smaller the difference between true and estimated relative partworths. We divide by the partworths by scale so that the relative importances sum to 1. Mathematically: $\left\| \vec{\beta}^{true}/\gamma^{true} - \vec{\beta}^{observed}/\gamma^{observed} \right\|$, where the $\vec{\beta}^{true}$ are the partworths that determine how the consumer reacts in the marketplace. Empirically, we test whether craft has an impact on accuracy, but our main theoretical focus is on observed scale.

### *Summary and Illustrations*

Figure 1 summarizes the relationships among accuracy, scale, internal validity, external validity, and the partworths. We parse the comparison of true partworths ($\vec{\beta}^{true}$) and the observed partworths ($\beta^{observed}$) into scale and accuracy. Both are affected by craft. Measures of marketplace choice enable scale adjustments to improve external validity. The next section links scale (and accuracy) to managerial decisions—managers act on observed partworths while the marketplace acts on the true partworths.

Using the conceptual model, we investigate how the quality of CBC studies (craft) affects observed scale and/or accuracy. For example, suppose that one attribute is described by realistic pictures and animations, but all other attributes are described by ambiguous text. We might expect that the enhanced description would enable consumers to better evaluate the realistically-described attribute and, perhaps, the better evaluations would increase the observed relative importance of that attribute compared to the consumer's true tradeoffs. This effect would reduce accuracy because the relative tradeoffs are now different from the consumer's true tradeoffs. On the other hand, if all attributes were described by realistic pictures and animations rather than ambiguous text, the relative partworths might not differ from the true relative partworths (accuracy), but the observed scale might differ from the true scale. (We later test this hypothesis.) In this case, the quality of the CBC study

affects observed scale but not accuracy. Other aspects of craft affect observed scale, accuracy, or both scale and accuracy. For example, Meissner, Oppewal, and Huber (2016) show that the number of alternatives in a choice set affects observed scale. Aribarg, Burson, and Larrick (2017) show that the measurement unit of attribute levels, e.g., days, hours, minutes, or seconds, affects relative partworths.

**Figure 1. CONCEPTUAL MODEL OF SCALE AND ACCURACY**

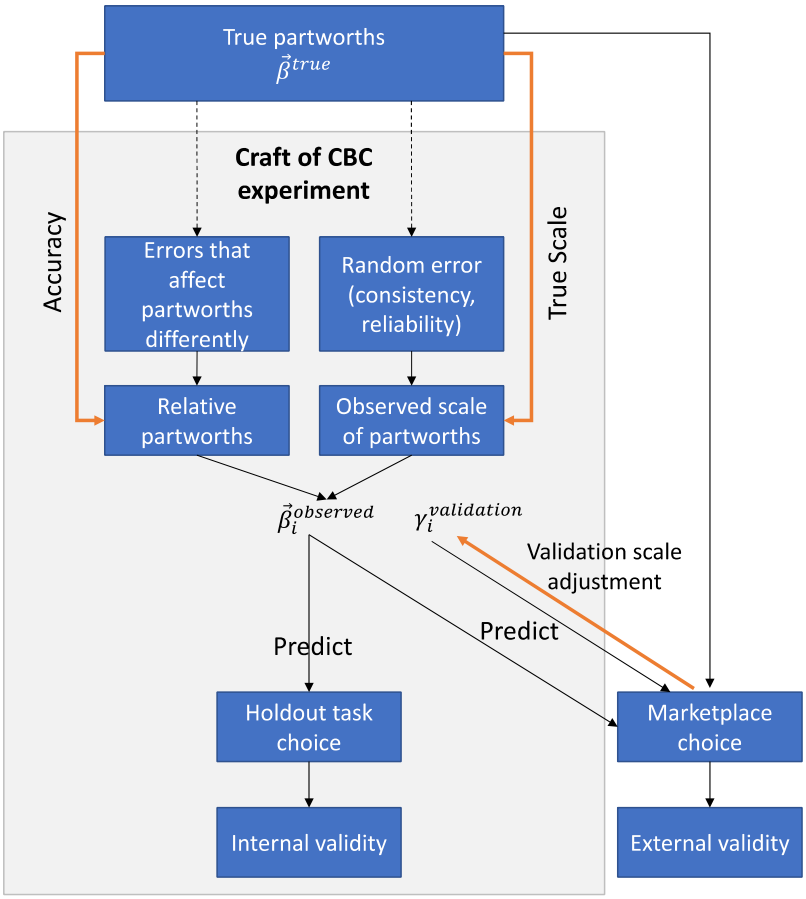



Table 1 ill ving three attributes of smar e the same true and observed partworths ($\beta_i = \beta, \eta_i = \eta, \gamma_i = \gamma$,) and that other attributes are not relevant or are held constant. We assume that currently only one (incumbent) smartwatch is on the market, i.e., a round, silver smartwatch with a black leather band that costs \$300 ($j = 1$). For illustration, we consider a new entrant ($j = 2$) that wants to differentiate from the incumbent and market a rectangular, gold smartwatch with a metal band for \$200. We focus only on consumers that are in the market to buy a smartwatch so that we can disregard the outside option (the theoretical model and the empirical

applications both consider the outside option).

Suppose the true relative partworths and true scale are given in the column marked "true partworths." For this column, the true utilities of the two products are $u_1^{true} = 3.0$ and $u_2^{true} = 1.0$. The probability that consumers choose $j = 2$, i.e., its expected market share, is $P_1^{true} = 0.12$. We temporarily ignore the fact that the incumbent might change its price to respond to the entrant—a phenomenon we address in the next section. Assuming a market size of 1 million consumers, initial investment costs of $5 million, and variable costs of $100 per smartwatch sold, the new entrant would earn a profit of $6.9 million and would be advised to launch their new smartwatch.

The column marked "lower accuracy" illustrates partworths with the same observed scale, $\gamma^{true}$, but the less-accurate partworths reverse the relative importances of watch face and price. Firms acting on lower accuracy might choose to improve the wrong attribute levels for a new product, say the watch face because consumers appear (incorrectly) to be relatively more sensitive to this attribute (e.g., Green, Krieger, and Wind 2001). In our illustrative example, the utility of the new entrant's product is biased downwards and the utility of the incumbent's product is biased upwards. The inaccurate model predicts $P_2^{lower\ accuracy} < 0.05$. Consequently, the new entrant would falsely predict a loss of $0.26 million and would not launch the new smartwatch. Because this phenomenon is well-known, we need not motivate further managerial interest in accuracy. However, the impact of scale is less well-known.

The column marked "lower-scale" illustrates partworths with the same relative values as the true partworths, i.e., high accuracy, but with partworths that are scaled uniformly lower. In this case, predictions of choice probabilities tend more towards chance. The lower-scale model predicts $P_2^{lower\ scale} = 0.27$ and profit predictions exceed $21 million. The new entrant will correctly introduce the product. However, the new entrant will realize profits that are substantially lower than predicted ($6.9 million vs. $21 million). The manager would likely be held accountable for not meeting his or her goal.

The last column represents a case where the firm is overly optimistic about scale, perhaps because it bases predictions on estimated partworths with a high consistency rather than validation-adjusted partworths. The overly-optimistic probability predicts that the first smartwatch ($j = 1$) will be chosen almost with certainty: $P_1^{optimistic} = 0.98$. Consequently, the new entrant would predict that it cannot achieve substantial market share (< 2%)and face losses of more than $3 million, The new entrant would incorrectly choose not to introduce a new smartwatch and, hence, miss out on the $6.9 million in profits they would have earned had they known the true scale. The specific numbers change when we take a price equilibrium into account, but the basic insight remains.

**Table 1. ILLUSTRATION OF SCALE AND ACCURACY**

| Illustrative Attributes of Smartwatches | True partworths | Lower accuracy, true scale | Lower scale | Overly optimistic scale |
|---|---|---|---|---|
| Round vs. rectangular face ($\gamma\beta_1$) | .50 | 1.00 | .25 | 1.00 |
| Silver vs. Gold-colored watch ($\gamma\beta_2$) | 1.00 | 1.00 | .50 | 2.00 |
| Black leather vs. metal band ($\gamma\beta_3$) | 1.50 | 1.50 | .75 | 3.00 |
| Price, \$200 vs. \$300 ($\gamma\eta$) | 1.00 | 0.50 | .50 | 2.00 |
| Observed Scale | 4.00 | 4.00 | 2.00 | 8.00 |

### *Heterogeneity*

For simplicity, Table 1 is based on homogeneous relative partworths and scale. However, empirically, scale and relative partworths vary across consumers. Modern CBC analysis uses estimation methods such as hierarchical Bayes or empirical Bayes to account for respondent heterogeneity (Allenby, et al. 2014a; Allenby and Rossi 1999; Lenk, et al. 1996; Sawtooth Software 2014). When partworths vary, we must account for heterogeneity. Specifically, we predict the market share of product $j$, $P_j$, by integrating over the distributions of scale values and relative partworths. If the distribution of scale and relative partworths and scale is $f(\gamma_i, \vec{\beta}_i)$, the percent of consumers who purchase product $j$ is given by Equation 3. Because some consumers may purchase the outside option, the shares over all products may add up to less than 100%.

(3) $$P_j = \int P_{ij} f(\gamma_i, \vec{\beta}_i) d\gamma_i d\vec{\beta}_i$$

Table 1 illustrated that both scale and accuracy affect predicted consumer response—a double whammy. These effects might be exacerbated if there is positive or negative correlation over respondents between scale and the relative partworths (e.g., Louviere 2001; Salisbury and Feinberg 2010)—a triple whammy.

## *SCALE AFFECTS MANAGERIAL DECISIONS FOR BOTH PRICES AND PRODUCT ATTRIBUTES*

### *Scale Affects Firms' Predictions of Market Price and Patent/Copyright Valuations*

Selecting the right price is clearly important to managers. In addition, the fastest growing application of CBC is in valuing patents and copyrights (Cameron, Cragg, and McFadden 2013). Indeed, CBC studies have become the norm rather than the exception in litigation studies. For example, a jury

awarded Apple, Inc. over $1 billion from Samsung, Inc. for damages due because Samsung infringed iPhone touchscreen patents. The damages were motivated, in part, by "price premiums" derived from a CBC study (Mintz 2012). A similar methodology was involved when the US Copyright Royalty Board set the royalty rates for streaming music (McFadden 2014). In both cases, craft was critical to whether or not the courts relied upon the CBC studies. The courts intuit that craft affects the projected value of a patent or copyright (e.g., Alsup 2012).

There are two approaches to make managerial recommendations on price and/or to value patents and copyrights. The first approach uses a CBC simulator to predict consumer choices in a hypothetical market in which a firm changes its price or product attributes. The resulting predictions are then interpreted by the managerial team (in the case of managerial recommendations) or a "damages expert" (in the case of patent/copyright valuations). The managerial team or damages expert might take other considerations into account (supply, marketing actions, competitive reaction) to make a recommendation or provide a value for the patent or copyright.

The second approach also uses a CBC simulator, but iterates that simulator to take competitive reaction into account within the simulations. A manager or damages expert might consider a change in one attribute and then use the CBC simulator to calculate a Nash equilibrium in prices. To calculate the Nash equilibrium, we define a profit function for each firm in the market and find the set of prices for all firms in the market such that each firm maximizes its profit and has no unilateral incentive to deviate from its equilibrium price. See computation methods in Allenby, et al. (2014a). A manager might choose the attribute levels that provide the largest equilibrium profit. A damages expert would compare the equilibrium profits obtained with the attribute level enabled by a patent to the equilibrium profits based on the best non-infringing attribute level.

Both approaches are sensitive to accuracy because, if the partworth of an attribute level relative to the price coefficient is inaccurate, the model will predict a greater (or lesser) response to price than the true relative partworths would predict. However, the second (equilibrium) approach is particularly sensitive to scale because a higher scale implies that choice probabilities, and hence profit, are more sensitive to differences in attribute levels or prices. The greater sensitivity implies, in turn, a market that is more sensitive to price and hence more competitive on price. The greater level of competition implies lower equilibrium prices.

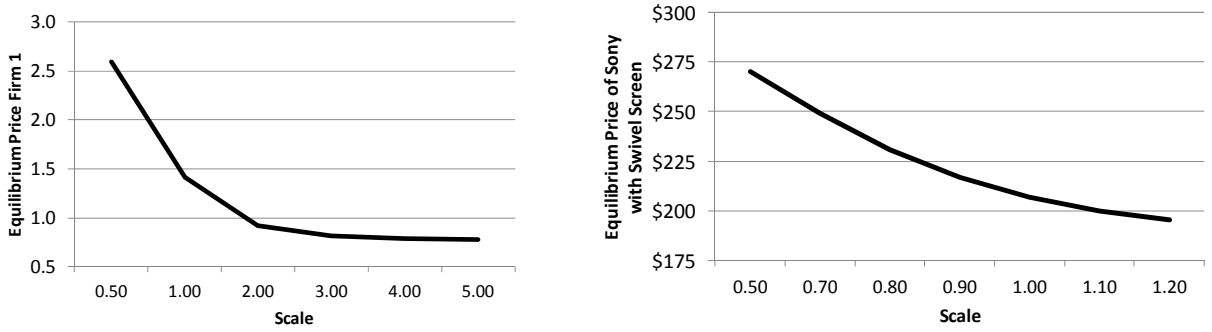
We use both a stylized model and empirical data to demonstrate that the predicted equilibrium prices depend upon the scale (observed, true, or validation adjusted) that is assumed when calculating equilibrium prices. Figure 2a plots the equilibrium prices from an analytic stylized model that isolates

the effect of scale. In that model, there are two consumer segments that vary on relative partworths, but share the same scale. Partworths are homogeneous within segment, but not between segments. For the analytical model, we simulate a duopoly in which one firm has the attribute level and the other does not. (Model details and parameter values are given in Anonymous 2017. The scale of price in Figure 2a is arbitrary.)

Figure 2b plots (counterfactual) equilibrium prices using the data and methods from the empirical study of cameras in Allenby, et al. (2014a). We provide details of the Allenby-et-al method when we analyze our own empirical data. For the empirical data, we use the firms and the attribute levels that are analyzed in Allenby, et al. We vary scale as a multiplier for all segments in the analytical model (Figure 2a) and for all consumers in the empirical data (Figure 2b).

**Figure 2. EQUILIBRIUM PRICE DEPENDS ON THE SCALE OBSERVED FOR THE CBC STUDY**

**a. Two-segment, Two-firm Stylized Model** **b. Four-firm Empirical Camera Market**




The results demonstrate that price equilibria are sensitive to scale and confirm that higher scale implies higher sensitivity, higher competitive response, and lower equilibrium prices. The results in Figure 2 are illustrative. We find similar price sensitivity for other attribute levels, other firms, and for the empirical data in this paper. Such differences in equilibrium prices are not trivial. A firm deciding to launch a new product with manufacturing costs of $200, will react differently if it expects the product will sell at $200 rather than $250. In patent/copyright cases, differences in "but-for" predicted prices could account for huge sums. For example, Apple claimed that Samsung infringed on over 20 million smartphones—even a $10 difference in damages per unit becomes $200 million in total damages.

### *Scale Affects a Firm's Strategic Selection of Product Attribute Levels*

Scale also affects firms' decisions on which attribute levels to include in new products. Suppose that the majority of consumers prefer silver-colored smartwatches relative to gold-colored smartwatches and suppose that the innovator is selling a silver-colored smartwatch. Consider a new

entrant to the market. If the marketplace is very sensitive to price (higher scale) and the new entrant (follower) also introduces a silver-colored smartwatch (and no other attribute levels vary), then competitive market forces will drive equilibrium prices down. The market will be less profitable for both firms. In this case, the follower may earn more profits if it introduces the less preferred gold-colored smartwatch to reduce price competition.

On the other hand, if the marketplace is not sensitive to price (low scale), the price equilibrium might be sufficiently high so that an undifferentiated market is more profitable for both firms (both firms offer the more-preferred silver-colored smartwatches). Anonymous (2017) formalize these arguments for heterogeneous CBC models and show that firms, acting optimally and in their own best interests and with knowledge of $\gamma^{true}$, choose to differentiate when scale is sufficiently high and choose not to differentiate when scale is sufficiently low. The result is similar in spirit to minimum-vs.-maximum differentiation arguments (e.g., d'Aspremont, Gabszewicz, and Thisse 1979, de Palma, et al. 1985; Economides 1984; Hotelling 1929), but in this case differentiation is driven by inherent stochasticity as measured by scale ($\gamma_i$'s) rather than heterogeneity in the relative partworths ($\vec{\beta}_i$'s) or by unobserved attributes. It is, perhaps, surprising that (true) scale alone is sufficient to drive a market from an undifferentiated market to a differentiated market (holding heterogeneity in the $\beta_i$'s constant). This is a different phenomenon from that typically analyzed in the differentiation literature. For example, de Palma, et al. (1985) use a logit-like function, but at the aggregate market level. Their aggregate error term represents heterogeneity in consumer preferences. In contrast, if heterogeneity in partworths is modeled explicitly—the magnitude of the error term represents inherent (unmodeled) stochasticity, not heterogeneity. The behavioral motivation underlying the effect of scale and the effect of heterogeneity is different and has different managerial implications.

The impact of scale on differentiation is important managerially. Suppose that a firm bases its attribute decisions on a CBC study that underestimates observed scale relative to true scale. It might choose not to differentiate its product. However, when the product is launched, the marketplace reacts according to the true scale. Suppose the true scale is higher such that differentiation would have been more profitable. The firm would have made a more profitable decision if it had observed a higher scale (perhaps with a higher-craft study) and differentiated.

We illustrate the phenomena by varying scale using the Allenby et-al. (2014a) empirical camera data. We compute prices and profits for a market in which Canon and Sony can add either Wi-Fi or a 10X optical zoom to their cameras. All other attribute levels are held constant and there are no other firms in the market. Consumers can choose to purchase from the innovator (here: Canon, $j = 1$), the follower

(here: Sony, $j = 2$), or not to purchase at all. Profit is sales (integration of logit-based probabilities) times margin (price minus variable costs).

Let $\pi^*_{1ww}$ be the equilibrium profits of the innovator if the innovator chooses Wi-Fi for its product (Canon, $j = 1$) and the follower chooses Wi-Fi for its product (Sony, $j = 2$). Define $\pi^*_{2ww}$ for the follower and define $\pi^*_{jzw}$, $\pi^*_{jwz}$, and $\pi^*_{jzz}$ similarly, for $j = 1, 2$, to indicate profits with either Wi-Fi (*w*) or zoom (*z*). We use the same CBC data and methods as Allenby, et al. (2014a), but we introduce a counterfactual scale factor to vary scale. For the purpose of this illustration, we assume that the marketplace reacts to the true scale and that both firms know the true scale.

We adopt the conventions of the literature and assume the innovator chooses its attribute level to optimize profit anticipating the reaction of the follower. Both the innovator and the follower assume that the market equilibrates to the Nash equilibrium prices. For example, in Table 2, when scale is lower (scale factor = 0.3), the innovator chooses zoom anticipating that the follower will also choose zoom because $\pi^*_{2zz} > \pi^*_{2zw}$. Had the innovator chosen Wi-Fi, the follower would have chosen zoom and the innovator would have been worse off. Because both the innovator and the follower choose zoom, the market is said to be undifferentiated. However, when scale is higher (scale factor 0.5), the innovator still chooses zoom, but the follower finds it in its best interests to choose Wi-Fi—a differentiated market. Table 2 demonstrates a general phenomenon. We find the same phenomena with our empirical data and for all attributes of cameras. Table 2 illustrates the directional effect—larger differences in scale imply larger differences in profits. In some cases, the differences in profits can be quite large.

**Table 2. TRUE SCALE AFFECTS THE NASH EQUILIBRIA IN PRODUCT ATTRIBUTE LEVELS**

| | **Lower Scale** | | | **Higher Scale** | |
|---|---|---|---|---|---|
| | **Follower: Wi-Fi** | **Follower: Zoom** | | **Follower: Wi-Fi** | **Follower: Zoom** |
| **Innovator: Wi-Fi** | $\pi^*_{1ww} = 71.1$<br>$\pi^*_{2ww} = 67.0$ | $\pi^*_{1wz} = 71.7$<br>$\pi^*_{2wz} = 69.6$ | **Innovator: Wi-Fi** | $\pi^*_{1ww} = 63.4$<br>$\pi^*_{2ww} = 59.6$ | $\pi^*_{1wz} = 66.3$<br>$\pi^*_{2wz} = 63.7$ |
| **Innovator: Zoom** | $\pi^*_{1zw} = 74.5$<br>$\pi^*_{2zw} = 67.7$ | $\boldsymbol{\pi^*_{1zz} = 73.9}$<br>$\boldsymbol{\pi^*_{2zz} = 69.0}$ | **Innovator: Zoom** | $\boldsymbol{\pi^*_{1zw} = 69.7}$<br>$\boldsymbol{\pi^*_{2zw} = 62.6}$ | $\pi^*_{1zz} = 67.1$<br>$\pi^*_{2zz} = 62.3$ |

### *Summary of the Effect of Scale on Managerial Decisions*

Scale in CBC is an overlooked phenomenon that has major implications for managerial decisions about pricing, about attribute levels for new products, and about patent and copyright valuations.

Knowing, or approximating, the true scale of the marketplace is critical when CBC is used to estimate market prices. However, managerial decisions depend upon the observed scale in the CBC study. We have argued that the quality of a CBC study (craft) affects observed scale such that higher craft leads to an observed scale that is closer to the true scale. Further, we argue that observed scale might depend upon whether it is based on estimated partworths or partworths adjusted with validation data. We now use empirical experiments to examine whether craft impacts observed scale sufficiently to affect managerial decisions and whether it matters if the firm bases its decisions on scale from estimation or as adjusted by validation. If we succeed in demonstrating that craft and validation matters, then we hope to encourage researchers to explore other aspects of a CBC study that might drive observed scale or other validation measures that mimic the marketplace better than internal holdout choice sets.

## *AN EMPIRICAL STUDY TO DEMONSTRATE THE EFFECTS OF CRAFT AND VALIDATION ON SCALE, ACCURACY, MANAGERIAL DECISIONS, AND PATENT/COPYRIGHT VALUATIONS*

To test whether craft drives observed scale, we choose four factors that the literature suggests are likely to affect the quality of a CBC study: (1) image realism, (2) incentive alignment, (3) training videos, and (4) instructions to hold all else equal (ceteris paribus condition). These aspects of quality are also cited often in litigation. Although some CBC studies use high levels of craft on one or more of these factors (e.g., Ding, et al. 2011; McFadden 2014; Mintz 2012), common practice relies on software defaults (lower levels of craft). We also test whether validation adjustments matter empirically and whether validation adjustments interact with craft in terms of recommended managerial actions.

### *Product Category and Sample*

To maintain high quality on those aspects of craft that we do not vary explicitly, we recruited 1,693 US respondents using a professional panel and followed recommended survey design principles (Schaeffer and Presser 2003). The panel, Peanut Labs, maintains 15 million pre-screened respondents. Peanut Labs is a member of the ARF, CASRO, ESOMAR, and the MRA and has won many awards for high quality. We screened for respondents who expressed interest in the category, were aged 20-69, and agreed to informed consent as required by our internal review boards. We pretested the questions, attributes, and choice tasks to assure they were easy to understand (66 pretest respondents). We tested for and found no demand artifacts—respondents were unaware and did not guess that the study was about craft or validation. We kept other potential aspects of quality constant across experimental conditions; we used sixteen choice sets for estimation with three profiles per choice set (plus one initial task for training, one holdout task, and one validation task). Three profiles is the most-common format in practice (Meissner, Oppewal, and Huber 2016). We used a dual-response format for the outside

option (Brazell, et al. 2006; Wlömert and Eggers 2016). (The dual-response choice task is illustrated in the next section. The detailed questionnaire is available from the authors.)

We chose a product category, smartwatches, which we believed would provide a reasonable test of whether the four aspects of craft affect observed scale. To keep the study feasible, we abstracted from the large number of attributes in smartwatches to focus on case color (silver or gold), watch face (round or rectangular), watch band (black leather, brown leather, or matching metal color), and price ($299 to $449). We were careful to assure that neither the text nor the images communicated unspecified attributes. Although unspecified attributes might affect the error term, and hence observed scale, this effect should be constant across image realism, incentive alignment, and training videos. To help isolate the effect of missing attributes, we manipulate whether or not respondents are told that all other attributes, including brand, are to be held constant (ceteris paribus condition). Because the brand is an important attribute of smartwatches, also because it is associated with the operating system, we hypothesized that such instructions (or lack thereof) would affect observed scale.

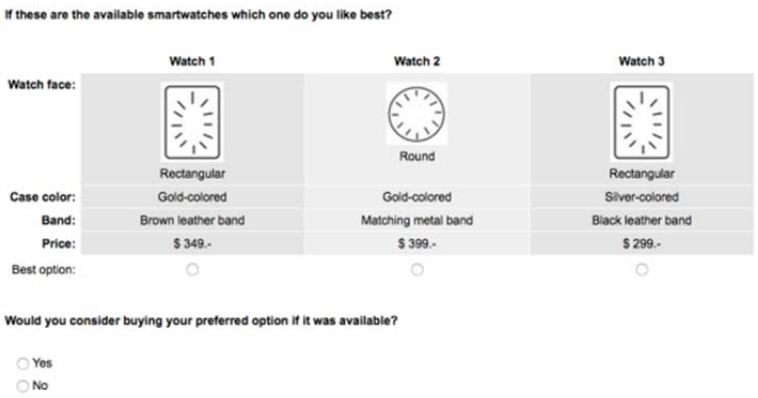
### *Four Aspects of Craft Varied in the Empirical Experiments*

We vary the realism of the stimuli, incentive alignment, the use of training videos, and the use of explicit ceteris paribus instructions in an orthogonal-array experiment: a half-replicate of a $2^4$ experimental design (eight orthogonal conditions). Respondents were assigned randomly to conditions.

***Realism of the stimuli.*** Past research suggests that rich visual representations are more realistic than text and more likely to evoke marketplace-like responses from respondents (e.g., Dahan and Hauser 2002; Vriens, et al. 1998). Dahan and Srinivasan (2000) suggest that visual profiles with animation might even be as effective as physical prototypes, while Dzyabura, Jagabathula, and Muller (2016) suggest that text-based profiles lead to different partworths than more-realistic products. Figure 3 gives example profiles for both the higher- and lower-realism stimuli. The higher-realism stimuli use realistic color graphics to represent the attributes (left side of Figure 3a). We animated the images such that a respondent could toggle among a detailed view, a top view, and an app view (static images illustrated in the right side of Figure 3a). The lower-realism stimuli used text plus limited graphics and represent the most common practice in CBC (Figure 3). Images could not be toggled in the lower-realism stimuli.

**Figure 3. HIGHER-REALISM AND LOWER-REALISM STIMULI**

(a) Higher-realism stimuli (left). Respondent could toggle the view (static examples on right)




(b) Lower-realism stimuli. Toggling was not available in this condition.

***Incentive alignment.*** We told respondents in the incentive alignment condition that 1 in 500 respondents would receive the smartwatch that the respondent chose in a randomly-selected choice task, plus the difference between the displayed price and $500 (Ding 2007; Ding, Grewal and Liechty 2005). The respondents would receive $500 in cash if the respondent chose the outside option. Because Wlömert and Eggers (2016) highlight that incentive alignment is often difficult for respondents to understand, we required incentive-aligned respondents to watch a one-minute video describing the incentives (e.g., https://youtu.be/DBLPfRJo2Ho). We adjusted the screens used in the videos to match the settings of the experimental conditions.

We told respondents in the non-incentive-aligned condition that 1 in 500 participants would receive a cash bonus of $500. Respondents who were not incentive-aligned were not shown an

incentive-alignment video. Pretests assured that respondents in the control condition did not believe that the $500 was tied to their answers.

***Training video.*** Although CBC is used widely, there is no guarantee that respondents understand the tasks or the attributes. Hauser, Dong, and Ding (2012) suggest that predictions based on CBC questions improve if respondents are well-trained. In the training-video condition, we required respondents to watch a two-minute animated video describing the smartwatch category, the smartwatch attributes, and the choice tasks (e.g., https://youtu.be/oji_bw_oxTU). The stimuli in each experimental condition were consistent with all other aspects of the experimental condition. For example, in the higher-realism image conditions, respondents saw a video that included higher-realism images. We modeled the training video on the videos that were used to study Samsung smartphones (Mintz 2012). We chose not to force an equivalent two-minute delay for respondents who were not assigned to the video-instruction manipulation. Such a delay would be unrealistic in the field and might create a demand artifact. To maintain realism and to test whether this particular set of video instructions implies higher craft, we purposefully tested the joint effect of training and fatigue, if any, due to watching the instruction video. This design enables us to determine empirically whether wear-out due to the two-minute training video offsets the advantage of improved respondent training.

***Ceteris paribus instructions.*** CBC formats limit the number of attributes that can be varied in a profile (Haaijer and Wedel 2007). As a result, best practice highlights to respondents that all other attributes are held constant in the choice tasks (Allenby, et al. 2014b, p. 642; ). We call these instructions ceteris paribus instructions. If respondents are not provided with ceteris paribus instructions, they might infer unobserved characteristics to be correlated with the attributes that are varied (Bradlow, Hu, and Ho 2004, p. 370). For example, without ceteris paribus instructions that highlight that other attribute levels do not change, respondents might be more likely to infer that quality changes if prices change. For respondents assigned to the ceteris paribus manipulation, we used the following phrase, repeated for every choice set.

> *Please assume that all watches are from your preferred brand [adjusted to consumer's brand preference] and are compatible with your smartphone so that they can show incoming messages or calls. Assume that all of these watches have a battery that lasts a day or more, a heart rate monitor, Bluetooth, high definition color LED touchscreen, 1.2 GHz processor, 4 GB storage, and 512 MB RAM*.

Respondents not assigned to the ceteris paribus condition, received no instructions and no reminder. As with the training video, we determine empirically whether any wear-out or information overload due to additional instructions offsets the advantage of the instructions.

### *Validation Tasks*

We used a holdout choice task to compute hit rates and uncertainty explained (internal validation). The holdout task was placed in the middle of the choice sets and was otherwise similar to those choice sets used for estimation; respondents were unaware these choice tasks were for internal validation.

External validation is a greater challenge. The ideal to which we strive is a test of whether the CBC model predicts the choices consumers would make if the hypothetical profiles became real products in the marketplace. Although we can get hints from the marketplace, e.g., does a silver-colored smartwatch sell better than a gold-colored smartwatch, marketplace sales are subject to many phenomena, such as advertising, promotion, distribution, and other marketing actions. Furthermore, many of the hypothetical profiles will never be market tested. Instead, we seek to mimic marketplace choices by creating a "market" that approximates the marketplace as closely as feasible while controlling for unmodeled marketing actions. As an initial demonstration of the importance of validation scale adjustment, it is sufficient that the validation choice task be perceived as closer to marketplace choices than the holdout task. If observed scale varies substantially between scale adjusted to the validation task (tested here) and scale based on the estimated partworths (typical practice), then we demonstrate that validation adjustment matters.

To mimic the marketplace, we created an incentive-aligned "market" in which respondents had a choice among smartwatches that they could purchase from a cash budget (or choose to receive the full cash amount). We made available as products every combination of the attributes levels. The task was substantively and visually different—it included all twelve, rather than three, products depicted by realistic images and used a different display format. To minimize short-term memory effects and to minimize respondent motivation to be internally consistent, the "market" occurred three weeks after the CBC data were collected. To determine if the time delay mattered, we also collected "market" data immediately following the CBC tasks.

External validation tests of CBC are rare in the literature. We feel our task is reasonably close to external validation. In a post-test, respondents found our marketplace task to be much closer to the true marketplace than holdout tasks ($p < 0.01$). Although the external validation task shares higher-realism images and incentive alignment with two experimental manipulations, we attempted to minimize

confounds with different formats and the time delay. At minimum, our marketplace validation task tests whether or not scale adjustments based a validation task affect managerial decisions relative to decisions using the observed scale based on the estimation. If validation adjustments matter managerially, then firms should be wary of not attempting external validation.

### *Descriptive Results—Time to Complete Tasks*

In total, 1,147 respondents completed both waves of the study (CBC task and the delayed validation task). To focus on respondents who were interested in smartwatches, we excluded 103 respondents who always chose the outside option. Neither the retention rate (wave 1 to wave 2) nor the exclusion varied significantly across experimental conditions ($\chi^2 = 6.72, p = 0.46$ for retention, $\chi^2 = 3.28, p = 0.86$ for exclusion). This leaves 1,044 respondents split randomly and approximately equally among the eight experimental conditions.

Because our experimental design is orthogonal, the manipulation of any one aspect of craft is not correlated with any of the other aspect of craft. To focus on the effect of each of the four manipulations, we present our results as contrasts between the two levels of craft for each manipulation. Results for each and every experimental condition are available from the authors. Table 3 summarizes the descriptive results.

**Table 3. DESCRIPTIVE RESULTS AS A FUNCTION OF CRAFT**

| | Image Realism | | Incentive Alignment | | Training Video | | Instruct Ceteris Paribus | |
|---|---|---|---|---|---|---|---|---|
| | Higher | Lower | Yes | No | Yes | No | Yes | No |
| Respondents exposed to manipulation | 518 | 526 | 521 | 523 | 521 | 523 | 522 | 522 |
| Median time to complete survey (seconds) | 476 | 476 | 505[a] | 436 | 541[a] | 399 | 482 | 466 |
| Median time to complete choice tasks (seconds) | 217 | 210 | 215 | 209 | 214 | 213 | 221 | 207 |

[a] significantly larger than alternative-level of craft at 0.05 level or better

The first row of Table 3 reports the number of respondents exposed to each aspect of craft. For example, 518 respondents saw higher-realism images while 526 respondents saw lower-realism images. Random assignment in experimental design means that roughly half of the 518 respondents and roughly half of the 526 respondents in these realism conditions were incentive-aligned, roughly half saw the training video, and roughly half received ceteris paribus instructions. The second row reports the total time to complete the CBC survey. As expected, the training video and the incentive alignment instructions increased the length of time the respondents spent on the study by 69 seconds (incentive-

alignment) and 142 seconds (training video) —basically the length of the videos. The third row reports the time to complete the choice tasks. None of the manipulations affected choice-set completion times.

## *ANALYSIS OF EMPIRICAL DATA*

Because our focus is on craft, scale, accuracy, and managerial decisions, we followed standard CBC practice to estimate models for each of the two experimental levels for each aspect of craft. We probed for interaction effects between the attributes and between aspects of craft. None of the interactions were significant; all results are reported for main-effects models. We interpreted the no-choice in the dual-response format as a choice of the outside option and we discarded the initial training choice sets that allow consumers to learn the CBC task. The posterior distributions of partworths were estimated with standard hierarchical Bayes software (in this case, the `bayesm` R package). To be consistent with Allenby, et al. (2014a), we assumed a normal prior distribution for all parameters, except for the price coefficient, for which we imposed a sign constraint by assuming a negative log-normal distribution. We drew 20,000 times from the posterior distribution, discarding the first 10,000 draws to achieve convergence, and keeping every 10$^{th}$ of the subsequent 10,000 draws. We perform our analyses for each of the 1,000 posterior draws. For exposition, we present the mean results across these draws.

### *Craft Affects Observed Scale and Predictive Ability*

In typical CBC applications, predictive ability is based on internal validity (holdout tasks). Researchers use a variety of predictive measures all of which are highly correlated on our data. For ease of exposition, we report the hit rate among the three smartwatches and the outside option and the percent of uncertainty explained ($U^2$, Hauser 1978). Table 4 suggests the (internal) predictive ability is comparable to that which we normally observe with a well-specified CBC model—the hit rates of 68-73% are well above chance (25%) and the models explain 39-47% of the uncertainty in the data. Higher-realism images and incentive alignment both improve predictive ability significantly relative to lower-realism images and no incentive alignment. This, too, is consistent with the literature cited earlier. The training video decreased predictive ability significantly. Ceteris paribus instructions had no effect.

We first calculate observed scale from the estimated partworths ($\gamma^{estimation}$) in order to assess the consistency of respondents' answers to the choice tasks. Craft does not necessarily enhance internal consistency. Based on the upper portion of Table 4, respondents are more consistent with text-based images (vs. higher-realism images) and with no incentive alignment (vs. incentive alignment), albeit the differences are not significant. The significant changes in consistency are reductions for the training video and the ceteris paribus instructions. The decrease in consistency and predictive ability due to the

training video suggests that the substantial time required to watch the training video likely wore out respondents causing a decrease that was more than the potential increase due to better knowledge of the smartwatch attributes and the choice task. Smartwatch attributes are well-known and the task is a natural choice task. Likewise, the lengthier ceteris paribus instructions might have increased respondent wear out leading to a lower internal consistency, albeit slightly. Ceteris paribus instructions are designed to increase external validity—a hypothesis we test next.

**Table 4. THE EFFECT OF CRAFT ON INTERNAL CONSISTENCY AND VALIDATION**

| | Image Realism | | Incentive Alignment | | Training Video | | Instruct Ceteris Paribus | |
|---|---|---|---|---|---|---|---|---|
| | Higher | Lower | Yes | No | Yes | No | Yes | No |
| *Holdout task, choice among three profiles plus outside option* | | | | | | | | |
| Hit rate (holdout task) | .73 [a] | .68 | .73 [a] | .68 | .68 [a] | .72 | .70 | .70 |
| Uncertainty explained ($U^2$, holdout task) [b] | .47 [a] | .39 | .46 [a] | .40 | .40 [a] | .45 | .43 | .42 |
| $\gamma^{estimation}$ [c] | 8.94 | 9.36 | 9.01 | 9.31 | 8.53 [a] | 9.42 | 8.86 [a] | 9.39 |
| *Validation task, delayed "market" choice among twelve products plus outside option* | | | | | | | | |
| Hit rate (validation) | .29 [a] | .21 | .27 [a] | .23 | .24 | .25 | .25 | .25 |
| Uncertainty explained ($U^2$, validation) [b] | .18 [a] | .10 | .16 [a] | .12 | .12 [a] | .15 | .14 | .15 |
| $\gamma^{validation}$ [d] | 4.11 [a] | 2.78 | 3.74 [a] | 3.12 | 3.20 [a] | 3.60 | 3.41 | 3.51 |
| Marginal impact of craft manipulations [e] | 1.48 | -- | 1.20 | -- | .89 | -- | .97 | -- |

[a] Significantly different than alternative level of craft at 0.05 level or better.
[b] Percent of uncertainty explained by model relative to that explainable by perfect prediction (Hauser 1978). Uncertainty explained is equivalent to relative Kullback-Leibler divergence (Ding, et al. 2011) for continuous probability predictions.
[c] Observed scale, calculated for each respondent, then averaged. Measures the ability to fit the estimation data.
[d] Validation-adjusted scale = observed scale from the estimation adjusted based on the validation task (logit model).
[e] Ratio of observed validation-adjusted scale comparing one level of craft to the other level of craft.

We now examine whether the effect of craft is different when predicting the delayed "market" validation task (lower portion of Table 4). Hit rate and $U^2$ for the validation task are naturally lower because they are based on predicting the outside option plus one of twelve market-based choices, rather than one of three holdout choices. Moreover, the validation hit rate and $U^2$ are based on a delayed task to minimize effects of learning. Hit rates are well above the chance level of 7.7% . $U^2$ is positive (more than chance) for all experimental cells. Higher-realism images and incentive alignment improve hit rates and $U^2$ significantly. Training videos lower $U^2$ significantly. Except for statistical ties,

the effect of craft on predictive ability is in the same direction for both the internal validity (holdout) and external validity (market) tasks.

We are most interested in the relative impact of craft on observed scale because observed scale has a substantial and direct effect on managerial decisions. We expect craft to affect validation-adjusted scale more than it affects estimation-based scale. (For example, respondents might consistently answer text-based choice sets, but their answers may not reflect how they would react in the marketplace.) To adjust observed scale for the validation task, we estimate an adjustment factor, $\alpha$, with a one-parameter logit model. The dependent measure is the market choice in the validation task. The explanatory variables are the estimated utilities, $\hat{u}_{ij} \equiv \hat{\beta}_i^{estimated}\vec{a}_j - \hat{\eta}_i^{estimated}p_j$ . (^ indicates estimated values.)

(4) $$P_{ij} = \frac{e^{\alpha\hat{u}_{ij}}}{e^{\alpha\hat{u}_{io}} + \sum_{k=1}^{J} e^{\alpha\hat{u}_{ij}}}$$

Validation-adjusted scale, $\gamma^{validation}$, is the adjustment factor times the observed scale based on the estimation ( $\gamma^{validation} = \alpha\hat{\gamma}^{estimation}$). We estimate adjustment factors for each level of craft. As a check, we also computed an adjustment factor using a joint likelihood to estimate the $\gamma_i$, the $\vec{\beta}_i$, and $\alpha$ simultaneously. There were some minor differences, but the two types of estimates were highly correlated ($\rho = 0.995$). Table 4 reports $U^2$ based on Equation 4. However, the directionality of the impact of craft on $U^2$ is identical without the $\alpha$ adjustment.

When we combine the adjustment factors ($\alpha$'s) with the $\hat{\gamma}^{estimation}$'s, we find that validation-adjusted scale is significantly larger for higher-realism (48% above lower-realism) images and incentive alignment (20% above no incentive alignment). The training video lowers validation-adjusted scale, mostly due to the significantly lower observed estimation-based scale ($\hat{\gamma}^{estimation}$). However, the magnitude of the effect is smaller (-11%) than that for realistic images and incentive alignment. Ceteris paribus instructions have no effect on validation-adjusted scale.

We examined whether the time delay of the "market" choice mattered. Both predictive ability and validation-adjusted scale were lower when based on the delayed "market" compared to the more-immediate "market." But, the realistic-image and incentive-alignment effects on $\alpha$ and on predictive ability were of the same relative magnitude and the same directionality for the immediate and delayed "market." No implications changed with respect to the craft manipulations. Details are available from the authors.

Summarizing the insights from Table 4, higher-realism images and incentive alignment are

clearly better craft. Both increase predictive ability and validation-adjusted scale substantially. The training video decreases validation-adjusted scale, but the effect is small with minor effects on predictive ability in the validation task. For our particular implementation of a training video, the wear-out effect appears to be slightly larger than the training effect. This result cautions practitioners that training videos must be designed to communicate information quickly and concisely.

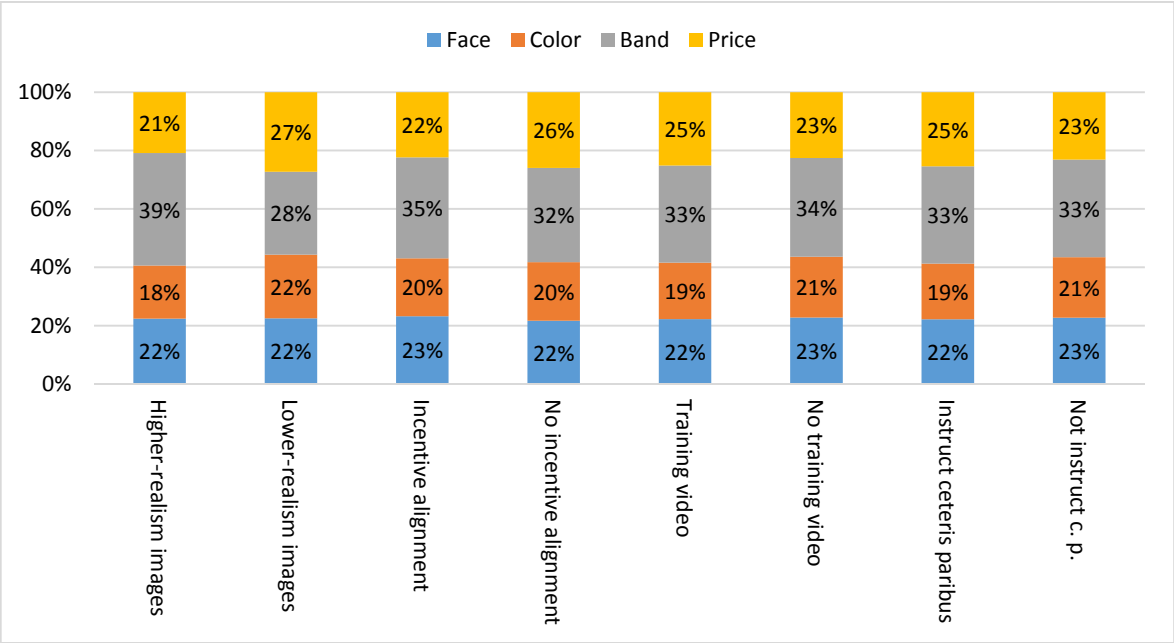
We obtain three interesting insights relative to CBC practice. First, the effect of higher-realism images is larger than the effect of incentive alignment as implemented in our experiments (last row of Table 4). Image realism has gotten far less attention in the CBC literature. Second, ceteris paribus instructions have little impact. While the effect of ceteris paribus instructions is rarely tested, the literature expects that such instructions are extremely important to external validity. Third, validation-adjusted scale is different from estimation-based scale. We show later in this paper that this difference has substantial managerial implications.

### *Impact on Accuracy (Relative Partworths)*

Before we examine the impact of observed scale on predicted price response, we examine whether or not craft also affects accuracy (relative partworths). We find the visual summary in Figure 4 useful, as a way to display the relative attribute importances, independent of observed scale. To obtain Figure 4, we calculate relative importances for each respondent and each draw and present the average. Image realism has the most substantial impact on relative importances. Color, band, and price are all significantly different between higher- and lower-realism images: band is more important (28% to 39%), color less important (22% to 18%), and price less important (27% to 21%). The other experimental manipulations have smaller effects, although all manipulations affect significantly the relative importance of price. In prior studies, incentive alignment sometimes, but not always, increases the importance of price (e.g., Ding 2007; Ding, Grewal, and Liechty 2005; Miller, et al. 2011). In our study incentive alignment lowers the relative importance of price; incentive alignment is best perceived as a method to encourage respondents to consider carefully all of the CBC attributes, not just price.

Together Table 4 and Figure 4 illustrate the double whammy of craft. The realism of the images affects both observed scale and accuracy substantially. We now explore how scale and accuracy affect managerial decisions.

**Figure 4. CRAFT AFFECTS RELATIVE ATTRIBUTE IMPORTANCE**




### *Impact on Equilibrium Prices*

Although choices in the marketplace depend upon the true partworths, firms make predictions (and act) on the partworths they observe in CBC studies. To illustrate how craft affects predicted equilibrium prices, we calculate Nash equilibrium prices in duopoly markets in which the firms compete with one another on different levels of the same attribute: either round vs. rectangular watch faces, gold vs. silver colors, black leather vs. brown leather bands, or black leather vs. matching metal bands. We assume that one firm offers a product with one of the two attribute levels and the other firm offers a product with the other level of the attribute. Each firm sets its price to maximize profit taking into account the other firm's price. When the Nash price equilibrium exists, we compute the equilibrium prices using the iterative procedure outlined in Allenby, et al. (2014a). They apply a root-finding method to identify the equilibria. For each market, we calculated equilibria for each of 1,000 draws using the validation-adjusted partworths. Because the largest price level in the CBC experimental design was $449, any prices above $449 extrapolate the CBC experiment beyond the range of its underlying data. To respect this constraint, we cap prices at an upper limit of $449.[1,2]

[1] Unconstrained extrapolation would unrealistically exploit a small percentage of consumers who have extremely low price sensitivity and would buy at unreasonably high prices. Allenby, et al. (2014a, p. 438) address the same issue by modifying the prior distributions on the partworths.

[2] Nash equilibria are not assured for mixed logit models (Aksoy-Pierson, Allon, and Federgruen 2013). In our data, respecting non-extrapolation solves the existence problem in the majority of cases.

The effect of craft is dramatic for image realism and incentive alignment (Table 5). Across attributes, on average, predictions based on CBC studies using higher-realism images imply predicted price differences that are about three times larger than predictions based on lower- realism images; incentive alignment implies predicted price differences that are twice as large than those from no incentive alignment. Training videos and ceteris paribus instructions have only minor impacts. We also computed price equilibria based on partworths not adjusted for validation. The relative effect of craft is similar, e.g., equilibrium prices predicted from a higher-realism study are substantially larger than those predicted from a lower-realism study and equilibrium prices are substantially larger for incentive alignment versus no incentive alignment.

**Table 5. CRAFT AFFECTS DIFFERENCES BETWEEN EQUILIBRIUM PRICES (VALIDATION PRECISION)**

| | Image Realism | | Incentive Alignment | | Training Video | | Instruct Ceteris Paribus | |
|---|---|---|---|---|---|---|---|---|
| | Higher | Lower | Yes | No | Yes | No | Yes | No |
| Round to rectangular watch face | $89 | $28 | $71 | $39 | $49 | $70 | $51 | $51 |
| Gold-colored to silver-colored case | $30 | $24 | $36 | $16 | $32 | $20 | $26 | $16 |
| Brown leather to black leather band | $72 | $32 | $72 | $33 | $51 | $62 | $47 | $47 |
| Metal to black leather band | $52 | -$4 | $34 | $17 | $17 | $42 | $17 | $33 |

The effect of validation-adjustment on predicted prices is also substantial. (Table available from the authors.) Consistent with the results in Figure 2, predicted equilibrium prices are, on average, 72% larger when based on (lower-scale) validation-adjusted partworths versus partworths not adjusted for validation. If firms make managerial decisions based on estimated partworths only, they would substantially underestimate equilibrium prices and the resulting profits. The effect is sufficiently large that it could affect a firm's GO/NO GO decision on new product introduction. When CBC is used for patent/copyright valuations, judgments range from hundreds of millions of dollars to a billion dollars (e.g., Mintz 2012). For such high-profile court cases, the effect in Table 5 and/or the effect of estimation-vs.-validation scale easily implies hundreds of millions of dollars differences in judgments.

### *Impact on Willingness to Pay (Price Premiums)—A Measure of Consumer Demand*

Price premiums, also called willingness to pay (WTP), indicate the price that consumers would be willing to pay for a change in the level of an attribute. WTP is used in product development to

evaluate attribute levels for inclusion in a new product. Most practitioners recognize that, while WTP is a factor in the price at which a product will sell in the marketplace, it is not the market equilibrium price (Orme 2010, p. 87).

We calculate WTP using the two-product simulation method in which one firm offers a product with a round watch face and another a product with a rectangular watch face. We adjust the price differences until the two products have equal shares. The effect of craft on WTP is similar to that for equilibrium prices. (Table available from the authors.) The WTP and validation-adjusted equilibrium prices are highly correlated ($\rho = 0.74$). The estimated WTP is 49% higher than estimated equilibrium prices, as is consistent with economic theory. Both the correlation and the relative values are reassuring as they suggest face validity. (The correlation between WTP and equilibrium prices based on estimated partworths is also high, $\rho = 0.93$).

### *Impact of Validation on Strategic Positioning Decisions*

Strategic positioning depends upon both scale and accuracy. (The firm makes its decisions based on the observed partworths; it does not observe true market prices until after the products are launched to the market.) For example, suppose that a rectangular watch face is only slightly preferred (on average) to a round watch face and that true scale is sufficiently low to imply an undifferentiated market. Then, strategic errors can result if (1) the observed scale is overestimated or (2) the relative partworths (round vs. rectangular relative to price) are overestimated. Either or both errors imply incorrectly that the firms should differentiate. Interestingly, firms might make the right decisions for the wrong reasons. For example, the firm might overestimate observed scale (which alone would lead to differentiation), but underestimate the relative partworths (which alone would lead to no differentiation). If the latter error is larger than the former error, the firm might choose not to differentiate. In our data, craft affects both observed scale and accuracy. Thus, it is possible that errors in both the observed scale and relative partworths could lead to the right strategic decision for the wrong reasons. (Naturally, a firm could not count on errors cancelling in exactly the right way—it is better to observe scale and relative partworths that are close to the true values.)

To identify optimal strategic positioning strategies, we compare predicted equilibrium profits in duopoly markets that are either differentiated or not differentiated. We then compare the two markets to determine the optimal equilibrium positioning strategy (as illustrated in Table 2). We summarize here the basic results. Equilibrium prices and profits are available from the authors. To examine the partial impact of observed scale, we use the relative partworths from the higher-realism images condition ($\hat{\beta}_i^{higher\ realism}$'s) and scale the relative partworths using either the observed scale ($\gamma^{estimation}$'s) or the

validation-adjusted scale ($\gamma^{validation\prime}$s) from Table 4.

Table 6 summarizes how optimal strategic decisions for round versus rectangular watch faces depend upon the observed scale. All estimation-based scale values predict that it would be optimal to differentiate. However, none of the validation-adjusted scale values come to the same conclusion. Table 6 suggests that the common practice of relying on estimated partworths, not adjusted for validation, leads to strategic positioning errors (at least for our data).

Table 6 focused on the impact of observed scale and validation adjustment. However, craft as well as validation, affects observed equilibrium prices. When we allow both the observed scale and observed relative partworths to vary as a function of craft, we find that craft also affects strategic positioning decisions, especially when scale is adjusted for validation. For example, when strategic positioning decisions are based on the study with a training video, the CBC analysis suggests that managers should not differentiate their product, but when strategic positioning decisions are based on the condition without a training video, the CBC analysis suggests differentiation. Details are available from the authors.

**Table 6. STRATEGIC POSITIONING DECISIONS DEPEND UPON OBSERVED SCALE: ESTIMATION-BASED SCALE VERSUS VALIDATION-ADJUSTED SCALE**
**(Only scale is varied. Relative partworths remain the same in every condition.)**

| | Image Realism | | Incentive Alignment | | Training Video | | Instruct Ceteris Paribus | |
|---|---|---|---|---|---|---|---|---|
| | Higher | Lower | Yes | No | Yes | No | Yes | No |
| Estimation-based scale | Diff [a] | Diff | Diff | Diff | Diff | Diff | Diff | Diff |
| Validation-adjusted scale (validation task) | No Diff [b] | No Diff | No Diff | No Diff | No Diff | No Diff | No Diff | No Diff |

[a] In equilibrium firms choose to differentiate – one firm offers a rectangular watch face, the other a round watch face.
[b] In equilibrium, firms choose not to differentiate—both offer a rectangular watch face.

### *Summary of the Impact of Craft and Validation*

Craft matters because it affects both observed scale and accuracy. Validation matters because it affects the observed scale on which the firm bases its decisions. Observed scale and accuracy, in turn, affect estimates of price equilibria, patent/copyright valuations, and strategic positioning. Firms that shirk on craft risk serious strategic mistakes for both pricing and positioning decisions. Firms (and litigation experts) that do not take the extra step of attempting to adjust for validation, risk forecasting errors. Such forecasting errors could lead to strategic errors in pricing and positioning and/or in the valuations of patents/copyrights. For example, if we revisit the Allenby, et al. (2014a) patent valuations

based on the prices in Figure 2b, predicted equilibrium price differences (infringing vs. non-infringing attribute level) vary from approximately $28 to $38—a difference that could imply hundreds of millions of dollars difference when converted to profit and multiplied times the sales of a popular infringing product. We demonstrated these potential strategic and patent/copyright validation errors with four key examples of craft and a validation measure that, hopefully, better mimics the marketplace than holdout tasks. The basic insights are more-general—craft and validation matter.

## *DISCUSSION AND FUTURE DIRECTIONS*

CBC studies affect thousands of managerial decisions every year. More recently, CBC has become a standard method to help value patents and copyrights. Despite this practical impact and despite a huge academic literature, most applications continue to use defaults and base predictions on estimated partworths. Our research challenges conventional practice by demonstrating that managerial decisions and patent/copyright valuations depend upon craft and validation adjustments. We further challenge conventional wisdom by highlighting the strategic importance of scale, which is as important as accuracy for managerial decisions. Managerial decisions are based on the observed scale perceived by managers to represent the market. Unfortunately, managerial decisions based on estimation-based scale may be different than those based on validation-adjusted scale. (Validation-adjustment is rare among both practitioners and academics.)

### *Discussion of Findings*

Our theories, and their implications, are reasonably general. Our empirical analyses are designed to be relevant, but illustrative. Our findings are robust. We find consistent results when applying different estimation techniques (e.g., using posterior draws, draws from the hyperparameters, or likelihood functions that jointly consider both estimation and validation data). General interpretations do not change if we use immediate validation rather than delayed validation (although we prefer the latter), if we adjust the delayed validation for test-retest reliability, or if we do or not adjust $U^2$ to account for $\alpha$. Moreover, scale is not affected by sample size. When using a random half of the sample for the estimation, the results remain consistent (the standard error for the estimated adjustment factor, $\alpha$, becomes larger but the mean $\alpha$ remains in the same magnitude).

We find that higher-realism images and incentive alignment matter—this should not be surprising, but it is surprising that realism might be more critical than incentive alignment and that both are more critical than training videos and ceteris paribus instructions. (Even online stores increasingly provide realistic images and even animations.) The large effect of realism is promising because it can be

achieved for highly innovative products that do not exist on the market (e.g., Dahan and Srinivasan 2000). For such products, incentive alignment is very costly, if not impossible. More importantly, the focus on scale (as well as accuracy) helps researchers understand and explain observed managerial impacts for a wide variety of craft decisions.

### *Future Research*

We hope to encourage researchers to explore further craft, validation adjustment, and their impact on observed scale and managerial decisions. Interesting craft effects might include the number of alternatives in a choice set, the number of choice sets, the selection of attributes and attribute levels, the use and number of brand-specific constants, assumptions about unmodeled marketing actions, questionnaire modality (in-person, computer, tablet, smartphone), panel quality, estimation method, adaptive questioning, experimental designs, and formatting issues such as dual-response formats. For example, Allenby, et al. (2014a) use an embedded outside option while Allenby, et al. (2014b) advocate a dual-response outside option. Wlömert and Eggers (2016) show that a dual-response outside option (and incentive alignment) provides better predictions using a delayed validation choice of actual service adoptions. Meissner, Oppewal, and Huber (2016) explore the impact of the number of alternatives in a choice set and find that the sum of estimated importances is 64% larger for choice sets with two products than for choice sets with five products. Hofstetter, et al. (2013) explore the effect of category experience and interest on WTP, thus suggesting a promising line of research to explore the effect of respondent characteristics on observed scale. Research on validation formats would be interesting, e.g., alternative “markets” and/or simulated stores, such as those used to forecast new product acceptance, might or might not provide improved validation adjustment.

We did not find strong effects for either training videos or ceteris paribus instructions, perhaps because their positive impact was counterbalanced by respondent wear out. We are not ready to give up on these scale drivers, but our research highlights the need for carefully crafting all such aspects of a CBC study.

## *REFERENCES*


Allenby, Greg M., Jeff Brazell, John R. Howell, and Peter E. Rossi (2014a), "Economic Valuation of Product Features," *Quantitative Marketing and Economics*, 12 (4), 421-456.

Allenby, Greg M., Jeff Brazell, John R. Howell, and Peter E. Rossi (2014b), "Valuation of Patented Product Features," *Journal of Law and Economics*, 57 (3) (August). 629-663.

Allenby, Greg M. and Peter E. Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89 (March–April), 57–78.

Alsup, William (2012), "Order Granting in Part and Denying in Part Google's Daubert Motion to Exclude Dr. Cockburn's Third Report," Oracle America, Inc. v. Google, Inc. C 10-03561, United States District Court, Northern District of California. March 13.

Anonymous (2017), "The Strategic Implications of Precision in Conjoint Analysis," Working Paper.

Aribarg, Anocha, Katherina A. Burson, and Richard P. Larrick (2017), "Tipping the Scale: The Role of Discriminability in Conjoint Analysis," *Journal of Marketing Research*, 55 (April), 279-292.

Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28 (September), 273–83.

Aksoy-Pierson Margaret, Gad Allon, and Awi Federgruen (2013), "Price Competition Under Mixed Multinomial Logit Demand Functions," *Management Science,* 59 (8), 1817-1835.

Ben-Akiva, Moshe and Steven R. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge, MA: The MIT Press.

Bradlow, Eric T., Ye Hu, and Teck-Hua Ho (2004), "A Learning-Based Model for Imputing Missing Levels in Partial Conjoint Profiles." *Journal of Marketing Research*, 41 (November), 369-381.

Brazell, Jeff D., Christopher G. Diener, Ekaterina Karniouchina, William L. Moore, Válerie Séverin, and Pierre-Francois Uldry (2006), "The No-Choice Option and Dual Response Choice Designs." *Marketing Letters*, 17 (4), 255–268.

Cameron, Lisa, Michael Cragg, and Daniel L. McFadden (2013), "The Role of Conjoint Surveys in Reasonable Royalty Cases," *Law360*, October 16.

Dahan, Ely and John R. Hauser (2002), "The Virtual Customer," *Journal of Product Innovation Management*, 19 (5) (September), 332-354.

Dahan, Ely and V. Srinivasan (2000), "The Predictive Power of Internet-based Product Concept Testing Using Visual Depiction and Animation," *Journal of Product Innovation Management*, 17 (March), 99-109.

d'Aspremont, Claude, Jean Jaskold Gabszewicz, and Jacques-François Thisse (1979), "On Hotelling's Stability in Competition," *Econometrica* 47 (5), (September),1145-1150.

de Palma, André, Victor Ginsburgh, Yorgos Y. Papageorgiou, and Jacques-François Thisse (1985), "The Principle of Minimum Differentiation Holds Under Sufficient Heterogeneity," *Econometrica* 53 (4), 767-781.

Ding, Min (2007), "An Incentive-Aligned Mechanism for Conjoint Analysis," *Journal of Marketing Research*, 54 (May), 214-223.

Ding, Min, Rajdeep Grewal, and John Liechty (2005), "Incentive-aligned Conjoint Analysis," *Journal of Marketing Research,* 42 (February), 67–82.

Ding, Min, John R. Hauser, Songting Dong, Daria Dzyabura, Zhilin Yang, Chenting Su, and Steven P. Gaskin (2011), "Unstructured Direct Elicitation of Decision Rules," *Journal of Marketing Research* 48 (February), 116-127.

Dzyabura, Daria, Srikanth Jagabathula, and Eitan Muller (2016), "Using Online Preference Measurement to Infer Offline Purchase Behavior," Working Paper, New York University, New York, NY.

Green, Paul E., Abba M. Krieger, and Yoram Wind (2001), "Thirty Years of Conjoint Analysis: Reflections and Prospects," *Interfaces*, 31 (3), (May-June), S56-S73.

Haaijer, Rinus and Michel Wedel, "Conjoint Choice Experiments: General Characteristics and Alternative Model Specifications," Chapter 11 in Anders Gustaffsson, Andreas Herrmann and Frank Huber, *Conjoint Measurement: Methods and Applications*, Springer: New York, 2007. Table 1, p. 207

Hauser, John R. (1978), "Testing the Accuracy, Usefulness and Significance of Probabilistic Models: An Information Theoretic Approach," *Operations Research*, 26 (3), (May-June), 406-421.

Hauser, John R., Songting Dong, and Min Ding (2014), " Self-Reflection and Articulated Consumer Preferences," *Journal of Product Innovation Management*, 31 (1) 17-32.

Hofstetter, Reto, Klaus M. Miller, Harley Krohmer, and Z. John Zhang (2013), "How Do Consumer Characteristics Affect the Bias in Measuring Willingness to Pay for Innovative Products?," *Journal of Product Innovation Management*, 30 (5), 1042-53.

Hotelling, Harold (1929), "Stability in Competition," *The Economic Journal,* 39, 41-57.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, and Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15 (2), 173–91.

Louviere, Jordan J. (2001), "What if Consumer Experiments Impact Variances as Well as Means?: Response Variability as a Behavioral Phenomenon," *Journal of Consumer Research*, 28 (3), 506–

511.

McFadden, Daniel L. (2014), "In the Matter of Determination of Rates and Terms for Digital Performance in Sound Recordings and Ephemeral Recordings (WEB IV)," Before the Copyright Royalty Board Library of Congress, Washington DC, Docket No. 14-CRB-0001-WR, October 6.

Meissner, Martin, Harmen Oppewal, and Joel Huber (2016), "How Many Options? Behavioral Responses to Two versus Five Alternatives per Choice," *Proceedings of the 19th Sawtooth Software Conference*, Park City, UT, September 26-20.

Miller, Klaus M., Reto Hofstetter, Harley Krohmer, and Z. John Zhang (2011), "How Should We Measure Consumers? Willingness to Pay? An Empirical Comparison of State-of-the-Art Approaches," *Journal of Marketing Research*, 48 (1), 172-184.

Mintz, Howard (2012), "2012: Apple Wins $1 Billion Victory Over Samsung," *San Jose Mercury News*, August 24.

Orme, Bryan K. (2009), *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research, 2E,* Madison WI: Research Publishers LLC.

Orme, Bryan K. (2009), "Fine-Tuning CBC and Adaptive CBC Questionnaires," Technical Report, Sawtooth Software, Inc., Sequim Washington.

Orme, Bryan K. (2016). "Results of the 2017 Sawtooth Software User Survey," https://www.sawtoothsoftware.com/about-us/news-and-events/news/1693-results-of-2016-sawtooth-software-user-survey.

Rao, Vithala R. (2014), *Applied Conjoint Analysis*, New York, NY: Springer.

Salisbury, Linda Court and Fred M. Feinberg (2010), "Alleviating the Constant Stochastic Variance Assumption in Decision Research: Theory, Measurement, and Experimental Test," *Marketing Science*, 29 (1), 1-17.

Sawtooth Software (2005), "Client Conjoint Simulator," https://sawtoothsoftware.com/download/techpap/ccsmanual.pdf

Sawtooth Software (2014), "Discover-CBC: How and Why It Differs from SSI Web's CBC Software," Technical Report, Sawtooth Software, Inc., Sequim Washington.

Sawtooth Software (2015), "What is Conjoint Analysis," http://www.sawtoothsoftware.com/products/conjoint-choice-analysis/conjoint-analysis-software.

Schaeffer, Nora Cate and Stanley Presser (2003), "The Science of Asking Questions," *Annual Review of Sociology*, 29, 65-88.

Thaler, Richard and Cass R. Sunstein (2009), *Nudge: Improving Decisions about Health, Wealth, and*

*Happiness*, New York, NY: Penguin Books.

Train, Kenneth E. (2006). *Discrete Choice Methods with Simulation*. New York, NY: Cambridge University Press.

Urban, Glen L. and John R. Hauser (2004), "'Listening-in' to find and explore new combinations of customer needs," *Journal of Marketing* 68 (April), 72-87.

Vriens, Marco, Gerard H. Loosschilder, Edward Rosbergen, and Dick R. Wittink (1998), "Verbal versus Realistic Pictorial Representations in Conjoint Analysis with Design Attributes," *Journal of Product Innovation Management,* 15 (5), (September), 455-467.

Wlömert, Nils and Felix Eggers (2016), "Predicting New Service Adoption With Conjoint Analysis: External Validity of BDM-Based Incentive-Aligned and Dual-Response Choice Designs," *Marketing Letters*, 27 (1), 195-210.