THE EFFECTS OF INCENTIVE ALIGNMENT, REALISTIC IMAGES, VIDEO INSTRUCTIONS, AND CETERIS PARIBUS INSTRUCTIONS ON WILLINGNESS TO PAY AND PRICE EQUILIBRIA

FELIX EGGERS

UNIVERSITY OF GRONINGEN
FACULTY OF ECONOMICS AND BUSINESS

JOHN R. HAUSER

MIT SLOAN SCHOOL OF MANAGEMENT
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MATTHEW SELOVE

GEORGE L. ARGYROS SCHOOL OF BUSINESS AND ECONOMICS
CHAPMAN UNIVERSITY

ABSTRACT

We describe how craft in conjoint analysis surveys (realistic images, incentive alignment, training videos, and ceteris paribus instructions) affects both accuracy (relative partworths) and precision (scale of the partworths). Accuracy and precision, in turn, affect estimations of willingness to pay (WTP) and predictions of market-equilibrium prices and profits for various "what-if" scenarios. Managerial recommendations, which attributes levels to include in a product and how to price a product, vary dramatically depending upon the craft of the study. When used in litigation, craft also affects the estimated value of copyrights and patents. To demonstrate the effect of craft, we conducted an experiment (smartwatch application) in which we systematically varied different drivers of craft. The use of realistic images increased accuracy and precision. Incentive alignment increased precision, but not accuracy. Neither training videos, nor ceteris paribus instructions had a positive effect. In fact, training videos reduced precision substantially because the wear-out effect (for our data) overwhelmed the training effect. The effect of craft on accuracy and precision had dramatic effects on estimations of WTP and equilibrium prices and profits. Managerial recommendations depended critically on craft as well as whether precision was based on the estimation data or adjusted for external validity. Craft matters! Defaults are not sufficient and could lead to incorrect recommendations and valuations.

MOTIVATION

Modern conjoint-analysis has been successful in the sense that it is now relatively easy to implement a conjoint-analysis study. For example, students using Sawtooth Software's Discover package can create conjoint-analysis designs and questionnaires within minutes with simple-to-understand point-and-click methods. Although pictures and animations are feasible, the default in most software packages is that profiles are described using plain text. Furthermore, advanced methods such as incentive alignment and instructional videos are difficult and expensive to implement. Not surprisingly, many conjoint studies rely on defaults. But does it matter? Does "craft" matter?

Everyday, both academics and practitioners make cost-vs.-benefits decisions about how to implement a conjoint-analysis study. Higher craft, e.g., more realistic pictures or animations, are often expensive and time-consuming. We would like to know whether the additional cost is justified. For example, are managerial recommendations sensitive to craft decisions such as the selection of images for products and attributes, the use of incentive alignment, the use of video instructions, enhanced instructions that "all else is equal," the use of dual-response formats, the number of alternatives in a choice set, the number of choice sets, the inclusion of attributes that describe the product in question but are not of managerial interest, etc.?

The question of craft is extremely important. Not only are there roughly 14,000 conjoint-analysis applications per year (Orme 2009, p. 127), but conjoint analysis is now being used to value copyrights and patents, often resulting in judgments in the hundreds of millions of dollars (Cameron, Cragg, and McFadden 2013). Through theory and empirical examples, we demonstrate that craft does matter. Craft affects pricing decisions, strategic decisions on which attributes to include in a new product, predictions of market response, predictions of profits due to managerial actions, and the valuations attached to copyrights and patents.

CRAFT AFFECTS BOTH ACCURACY AND PRECISION

For the purposes of this paper, we distinguish two aspects of a conjoint-analysis study that might be affected by craft: accuracy and precision. We call a conjoint-analysis study *accurate* if it estimates the correct relative partworths. We call a conjoint-analysis study *precise* if it estimates the correct scale, that is, the correct absolute magnitudes of the partworths. Precision measures the signal-to-noise ratio, because it compares that which is explained by the attributes in the conjoint design to that which remains noise (the error term).

To define accuracy and precision mathematically, consider the standard logit model that is used in most choice-based conjoint analyses. For ease of exposition, we write the equation for binary attributes and for dummy-variable coding. The concepts apply to effects coding and to multi-level attributes. Indeed, our empirical example includes a multi-level attribute.

Let u_{ij} be consumer i's utility for product profile j. Let δ_{jk} be a binary indicator for the k^{th} attribute such that $\delta_{jk} = 1$ if attribute k is at its "higher" level for profile j and $\delta_{jk} = 0$ if attribute k is at its "lower" level for profile j. Although we say "higher" and "lower," we do not require that the higher level be preferred to the lower level, or that the ordering of levels is the same across consumers. For example, the "higher" level might be a silver-colored product and the "lower" level might be a gold-colored product. Let p_j be the price associated with the j^{th} product profile.

To fully specify the utility function, we define β'_{ik} as consumer i's (raw) partworth for attribute k ("higher" vs. "lower" level), η_i as the weight for price, and ϵ_{ij} as an i.i.d. Gumbel-distributed error term. Assume there are K attributes. Utility for the j^{th} product profile is specified as:

$$u_{ij} = \sum_{k=1}^{K} \delta_{jk}\beta'_{ik} - \eta_i p_j + \epsilon_{ij} \tag{1}$$

The logit model predicts the probability, P_{ij}, that consumer i chooses the j^{th} product profile for a choice set consisting of J product profiles. In this equation, we let u'_{io} denote the utility of the no-choice option.

$$P_{ij} = \frac{e^{\sum_{k=1}^{K} \delta_{jk}\beta'_{ik} - \eta_i p_j}}{\sum_{\ell=1}^{J} e^{\sum_{k=1}^{K} \delta_{\ell k}\beta'_{ik} - \eta_i p_\ell} + e^{u'_{io}}} \tag{2}$$

Following McFadden (2014), we rescale utility to include a scale factor, γ_i, such that the relative weight on price is 1.0. In this formulation, as interpreted by McFadden (2014), the β_{ik}'s are the amounts that respondent i is willing to pay (WTP) for moving attribute k from its "lower" level to its "higher" level. (If the lower level is preferred, the WTP is negative.) Note that the WTP does not depend upon γ_i.

$$P_{ij} = \frac{e^{\gamma_i(\sum_{k=1}^{K} \delta_{jk}\beta_{ik} - p_j)}}{\sum_{\ell=1}^{J} e^{\gamma_i(\sum_{k=1}^{K} \delta_{\ell k}\beta_{ik} - p_\ell)} + e^{\gamma_i u_{io}}} \tag{3}$$

We call γ_i the precision (for consumer i). In the conjoint analysis literature, γ_i is sometimes called the "scale factor." The basic concept is that if γ_i is large, then the standard deviation of the error term is small compared to the magnitude of the partworths. A small relative error term means that the CBC logit model predicts more precisely the consumer's choices. (We say "relative" because, in most logit specifications, there are K parameters for K partworths. Thus the standard deviation of the error term is not identified independently of the partworths—only the relative magnitude of the error term is identified.)

We illustrate accuracy and precision with Table 1. Consider a conjoint design with three binary attributes and price. Suppose that the true raw partworths represent how consumers actually behave in the marketplace when making choices among products described by these three attributes and price. (The error term includes all unmodeled effects including attributes that are not accounted for and any inherent uncertainty in consumer choice.) These partworths are shown in the first column of data in Table 1.

Table 1. Illustration of Precision and Accuracy

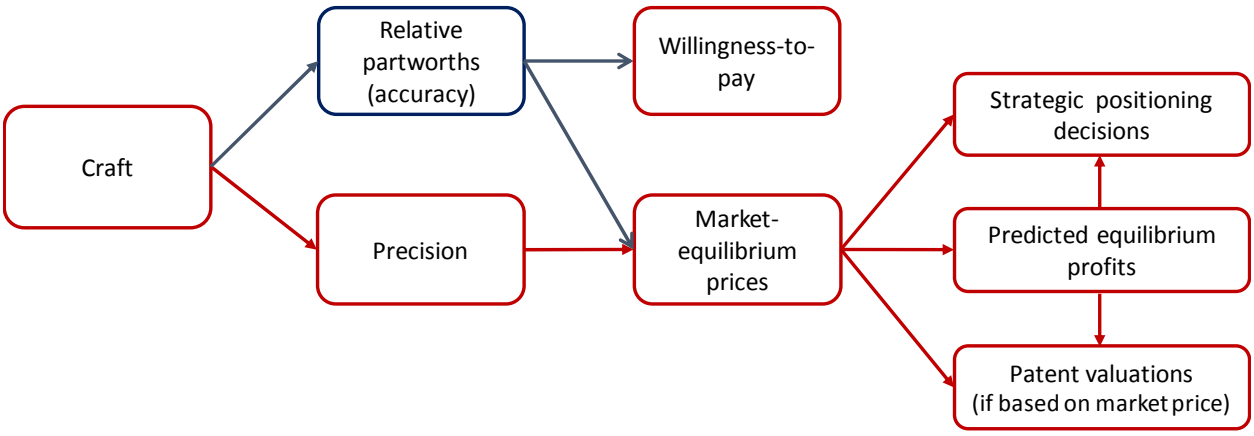
	True (raw) partworths	Lower accuracy	Lower precision	Higher Precision
Attribute 1	1.0	2.0	0.50	2.0
Attribute 2	2.0	1.0	1.00	4.0
Attribute 3	0.5	0.5	0.25	1.0
Price	1.0	1.0	0.50	2.0

Suppose that we estimate a model that has roughly the same magnitude of partworths, but it switches the importances of attributes 1 and 2. We say that such a model has lower accuracy (2nd column of data). Suppose we estimate a model that gets all relative partworths correct, but has a lower scale. We say that such a model is accurate but less precise (3rd column of data). Finally, the last column of data illustrates a model that is accurate but appears (to the analyst) more precise than truth.

Equation 3 is a useful theoretical definition of precision, but, in a population of consumers we might prefer a definition of precision that takes into account the fact that both the relative partworths (β_{ik}'s) and the precisions (γ_i's) vary over respondents. The ability to model such variation is an important advantage of advanced models such as hierarchical Bayes or empirical Bayes. When partworths vary or when accuracy differs between studies, a better empirical measure of the precision is the average of the respondents' sum of absolute attribute importances (Arora and Huber 2001). We use that measure in our empirical comparisons throughout this paper.

Having defined accuracy (relative partworths) and precision (scale factor), we now hypothesize that craft affects both and we hypothesize that both accuracy and precision affect managerial recommendations. We illustrate these hypotheses in Figure 1.

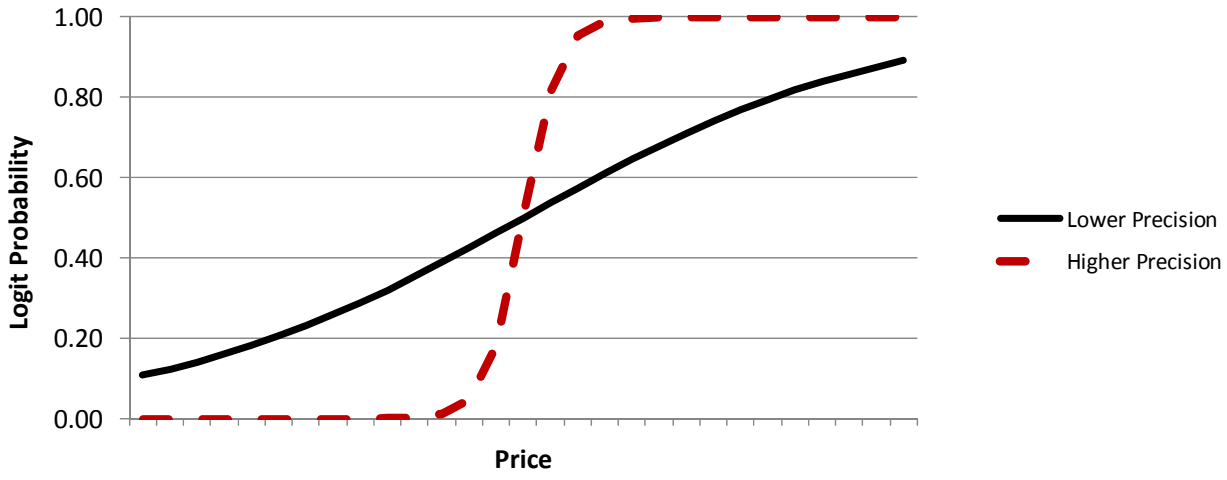
Figure 1. Hypotheses: Craft Affects Managerial Recommendations



PRECISION AFFECTS PRICE SENSITIVITY, WHICH, IN TURN, AFFECTS PREDICTED MARKET-EQUILIBRIUM PRICES AND THE STRATEGIC SELECTION OF ATTRIBUTES FOR PRODUCTS

When true precision is higher, consumer choices are more sensitive to changes in attributes or prices. We illustrate this phenomena in Figure 2 where we use Excel to plot the logit probabilities, P_{ij}, as a function of prices (p_j) for lower precision (γ^{lower}) and for higher precision (γ^{higher}). When precision is lower, as in the solid line in Figure 2, choice probabilities change more slowly as price changes. The curve is much flatter, almost a straight line. However, when precision is higher, as in the dotted line in Figure 2, choice probabilities change more quickly as price changes. The curve is much steeper. For sufficiently high precision ($\gamma_i \to \infty$), the logit model acts like a first-choice model (a step function).

Figure 2. Higher Precision Means Greater Price Sensitivity



WILLINGNESS TO PAY. WTP clearly depends upon accuracy. This is clear in McFadden's (2014) formulation because consumer i's WTP for attribute k is the relative partworth (β_{ik}). If the β_{ik}'s

are incorrect, then the estimate of WTP will be incorrect. While WTP is not market price (Orme 2009, p. 87), WTP is valuable in its own right. For example, when Polaroid launched the iZone camera it was able to determine that consumers would pay, on average, close to $10 for interchangeable camera covers—a feature that cost but pennies to produce. On the other hand, it learned that consumers were unwilling, on average, to pay anywhere near the cost necessary for the iZone camera to produce higher-resolution photographs. (The iZone camera was targeted to kids and produced postage-stamp-sized instant pictures.) Polaroid included interchangeable camera covers, but not higher-resolution capability (McArdle 2000). Similarly, when valuing patents, the marketing expert is often asked to provide WTP estimates to "damages" experts who combine secondary data with WTP to arrive at valuations (McFadden 2014; Mintz 2012).

MARKET EQUILIBRIUM-PRICE. After a conjoint analysis model is estimated, the relative partworths and precisions can be used in a choice simulator. Choice simulators predict how aggregates of consumers (the market and/or market segments) react to changes in attributes or price. Allenby, et al. (2014) propose that conjoint-analysis market simulations be used to compute Nash equilibrium prices. They further propose that the courts rely on the marketing expert to be both a marketing expert and a damages expert by estimating the change in Nash equilibrium prices due to a patent. They propose that the output of the conjoint analysis simulator for a product with the patented feature be compared to the output of the conjoint analysis simulator for a product without the patented feature.

The same methods, as those proposed by Allenby, et al., can be used to predict how the market will react to the introduction of a new product or to a change in a product's attribute levels. If costs are known, the simulator can predict profits for the new product or for a change in a product's attribute levels. The analysis could be extended to situations where competitors are hypothesized to respond to a new product by changing their attribute levels. For example, if BMW introduces adaptive cruise control, we might expect Audi, Mercedes, and Lexus to respond by introducing their own versions of adaptive cruise control. (By the way, this has already happened.)

As illustrated in Figure 2 (for price), predictions of consumer price response depend upon precision. This sensitivity to precision is particularly critical if the simulator is used to predict equilibrium prices. For example, Hauser, Eggers, and Selove (2016) illustrate the sensitivity of equilibrium prices to precision with a simple two-segment model in which the relative partworths and precisions are homogeneous within segment (but not between segments). They then vary precision and compute the Nash equilibrium prices. These prices are shown in Table 2. Notice that equilibrium prices vary dramatically over the range of precisions that we might expect in empirical conjoint analysis studies.

Table 2. Precision Affects Equilibrium Prices

Precision (γ)	Predicted Equilibrium Price in Differentiated Market (in currency units)
0.5	2.59
1.0	1.42
2.0	0.92
3.0	0.82
4.0	0.79
5.0	0.78

In Table 2, the equilibrium prices vary from under 1.0 currency unit to over 2.5 currency units. This could have a dramatic effect on whether or not a product is launched or in the valuation of a copyright or patent. In a typical patent case, such a difference in equilibrium prices (with and without a patented feature) could mean a difference in valuation in the hundreds of millions, or even billions, of dollars. For example, Apple has sold over 800 million iPhones. If the difference in price due to a patented feature swung from $10 to $25 due to precision, that would imply a difference in valuation of $12 billion. (These prices are the equilibrium prices, not the differences in equilibrium prices. Estimating differences requires additional simulations, but the point is made. Prices depend dramatically on precision.)

If craft affects precision, then clearly craft matters for predicting price and profits, whether it be for a newly designed product, a change in a product's attribute levels, or due to a patented or copyrighted feature. It is, of course, obvious that accuracy affects WTP and hence also affects predicted equilibrium prices and profits.

STRATEGIC RECOMMENDATIONS FOR THE ATTRIBUTE LEVELS OF A PRODUCT. Suppose that more consumers prefer a silver-colored watch to a gold-colored watch than vice versa. An innovator of smartwatches, facing no competition (and limited to one color) might introduce a silver-colored smartwatch. Even if the innovator is anticipating a competitor, the innovator might try to preempt competition and "position" its product as a silver-colored smartwatch. In practice, such strategic positioning decisions are usually made on product "positions" that are difficult to match. For example, Brita filters preempted competition by positioning themselves as the best-tasting water filters. It's competitor, P&G's PUR water filter, differentiated the market by positioning itself as healthy. The market was also differentiated on attributes, with Brita dominating pitcher filters and PUR dominating faucet filters.

A follower now has a choice. If the follower ignores the fact that the innovator is marketing a silver-colored smartwatch, the follower might be tempted to introduce a silver-colored smartwatch. For example, the follower might use a conjoint analysis simulator without taking into account that the innovator will respond to the follower's launch. For example, the innovator might lower its price to combat the market entry. A more sophisticated follower might decide to differentiate the market and introduce a gold-colored smartwatch. The sophisticated follower

hopes that it will sell to the gold-color-preference segment leaving the silver-color-preference segment to the innovator, thereby reducing price competition.

The decision depends on the size of the two segments, on the market response to color, and on the market response to price. If the market is not very responsive to either price or color, differentiation will have little effect on reducing price competition. The follower might be best advised to target the largest segment and introduce a silver-colored smartwatch. On the other hand, if the market is very responsive to price or color, differentiation will reduce price competition. The follower might be best advised to introduce a gold-colored smartwatch.

Market response to both attributes and price depends upon precision. Hauser, Eggers, and Selove (2016) prove formally that lower precision implies an undifferentiated market, while higher precision implies a differentiated market. They also demonstrate that the intuition based on the formal analysis applies when heterogeneity is continuous as is modeled in standard conjoint analysis estimation. (While this result has the flavor of standard analyses of differentiation, the effect is due to precision not to heterogeneity in consumer preferences.)

Put another way, the strategic choice of which level to include in a product depends upon precision, even if the relative partworths are perfectly accurate. Of course, accuracy also affects the relative partworths and, hence, the choice of attribute levels for the firm's product. Craft matters for the strategic choice of product attributes!

TRUE PRECISION VERSUS THE ESTIMATED PRECISION UPON WHICH THE FIRM RELIES

Before we examine empirically four elements of craft (drivers of precision), we pause to discuss interpretations of precision. Typically, in conjoint analysis applications, analysts estimate a logit model and report the partworths. The absolute partworths might be used in a simulation, as is common in academic studies. Alternatively, the relative partworths might be used, combined with an analyst-chosen error magnitude, as in randomized first-choice simulations. In either case, there is a precision (γ) associated with the simulation. When this precision is based on the estimated logit model, that is, the choice sets used to estimate the logit model, we might call it estimated precision, $\gamma^{estimation}$.

Typically, analysts report fit statistics such as hit rate, the percent of uncertainty explained (U^2), Kullback-Leibler convergence, root likelihood, AIC, BIC, or DIC. Sophisticated analysts also report these statistics for holdout choice sets. Experienced analysts know that the holdout statistics are never as good as the fit statistics. The precision reported based on the fit data may not be the precision that applies to the holdout data. It is a simple matter to re-estimate precision in a one-variable logit model. The dependent variable in this logit model is choices in the holdout sets and the explanatory variable is utility as created by the relative partworths. We might call this data-based precision, $\gamma^{internally\ adjusted}$. We use the word "internally" because holdout choice sets are really a test of internal validity. For high-craft studies, we might expect that

internal validity is a good indicator of external validity, but that is worth testing. For one test, see Toubia, et al. (2003).

Although rare, some analysts go a step further and test a form of external validity. For example, the analyst might create an incentive-aligned marketplace and ask respondents to make choices in that marketplace after a delay of a few weeks. The closer the induced marketplace is to the real marketplace, the closer the test is to a test of external validity. We adjust precision for external validity with the same type of one-variable logit. The only difference is that the dependent variable is now the choice in the induced marketplace. We might call this externally-adjusted precision, $\gamma^{externally\ adjusted}$.

We hypothesize that $\gamma^{internally\ adjusted}$ and $\gamma^{externally\ adjusted}$ depend upon craft. That is, we expect that higher craft leads the analyst to estimate models that fit the holdout data better and fit any induced marketplace data better. We expect that the precision from the highest craft study, adjusted to the induced marketplace may be our best estimate of true precision, γ^{true}. Estimation precision, $\gamma^{estimation}$, may or may not depend upon craft. If consumers are consistent in their answers to lower-craft questions, $\gamma^{estimation}$ might be high even if we cannot predict holdout choices or choices in an induced marketplace.

In this paper, we report $\gamma^{estimation}$, because this is the most common precision that is reported (when precision is reported). We also report $\gamma^{externally\ adjusted}$ based on an incentive-aligned induced marketplace with twelve smartwatches and an outside option that takes place three weeks after the estimation (and holdout data) are collected. (One in 500 respondents received the smartwatch they chose in the induced market, plus any change from $500.) We take on faith that $\gamma^{externally\ adjusted}$ is closer to γ^{true} than is $\gamma^{estimation}$ or $\gamma^{internally\ adjusted}$.

AN EMPIRICAL STUDY TO TEST DRIVERS OF PRECISION AND ACCURACY

We chose to test four drivers of precision and accuracy: (1) relative realism of the images used to describe attributes and profiles, (2) incentive alignment, (3) videos that train respondents about attributes and the choice tasks, and (4) instructions that all attributes, not displayed explicitly in the product profiles, are to be considered common to all profiles in the choice set (ceteris paribus instructions). These four drivers are cited often in the literature and in challenges to the use of CBC to value patents and copyrights. If we can show that any of these elements of craft affect managerial recommendations, then we hope to encourage other researchers to examine additional elements of craft.

The context of the empirical test is smartwatches. This category is sufficiently complex to make the test relevant, but not so complex as to make an initial test infeasible. We focused on just three attributes and price: 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). Naturally, any missing features affect the error term, and hence precision, thus, γ^{true} remains finite. The effect of non-modeled attributes should be constant across all but the ceteris paribus

manipulation. In the ceteris paribus instructions, we inform consumers that brand, operating system, and technical features remain constant across all profiles. (Consumers are asked about their brand choice and all profiles are about that brand.) Because brand (and operating system) are important features of smartwatches, we hypothesized that ceteris paribus instructions (versus lack thereof) would affect precision. Holding brand and operating system constant means the ceteris paribus manipulation provides a strong test for the effect of non-modeled attributes in the smartwatch category.

We wanted to maintain high quality on all aspects of the study that were not experimentally varied. Such recommendations are made in Allenby, et al. (2014). We recruited a US sample from a professional panel. The panel, Peanut Labs, maintains 15 million prescreened respondents. Peanut Labs is a member of the ARF, CASRO, ESOMAR, and the MRA and has won many awards for high quality. We targeted respondents who expressed interest in the category, were aged 20-69, and agreed to informed consent as required by our internal review boards.

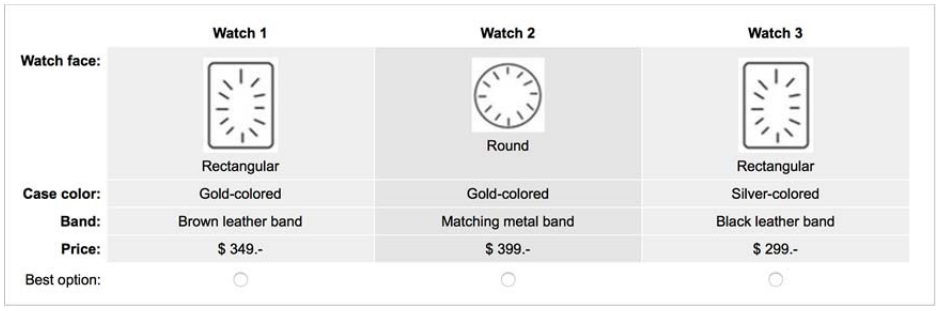
We followed standard survey design principles (Schaeffer and Presser 2003). All questions, attributes, and choice tasks were pretested carefully (66 respondents using the same sampling criteria). We tested for and found no demand artifacts. We used hierarchical Bayes logit-based estimation (Sawtooth Software 2009). We used sixteen choice sets for estimation (and two for holdout validation) with three profiles per choice set. (Three profiles is, by far, the most common in applications. Meissner, Oppewal, and Huber 2016.) We used a dual-response format for the outside option (Brazell, et al. 2006; Wlömert and Eggers 2016).

We varied the four elements of craft in an orthogonal half replicate of a 2^4 design. Respondents were assigned randomly to each experimental cell of the half replicate, with roughly an equal number of respondents in each cell. Three weeks after the conjoint-analysis studies were completed, we recontacted respondents and asked them to choose a smartwatch from a (full factorial) market of twelve smartwatches in a dual-response format. We assigned prices to the twelve smartwatches randomly and confirmed, based on the pretest data, that none of the options is dominating or dominated. The choice was incentive-aligned and presented the most realistic images of smartwatches.

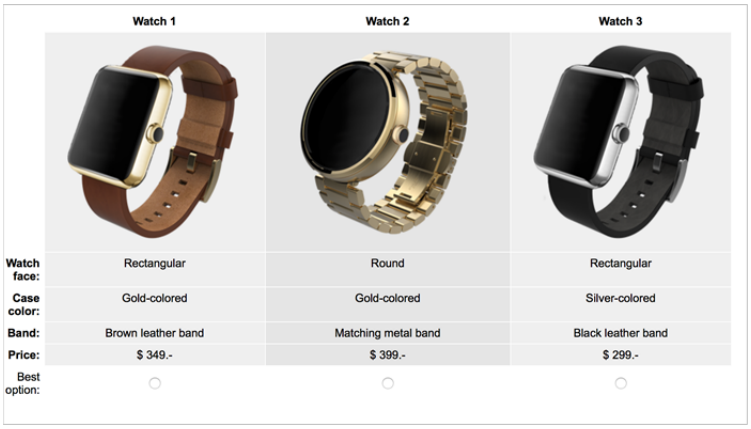
The four elements of craft were:

REALISM OF THE STIMULI. For the (hypothesized) higher level of craft we used realistic images of smartwatches in which the attributes were embedded in the images and listed in easy-to-read text below the images. The images included animations that allowed the respondents to see multiple views of the watches. For the (hypothesized) lower level of craft, we used a format similar to that which is the default in most conjoint-analysis software. The lower-craft profiles are primarily text-based, although we did use simple graphics. Figure 3 gives examples of the stimuli (but not the animations).

Figure 3. Varying the Craft of the Images (from Eggers, Hauser, and Selove 2016) (lower craft shown first, then higher craft)



	Watch 1	Watch 2	Watch 3
Watch face:	Rectangular	Round	Rectangular
Case color:	Gold-colored	Gold-colored	Silver-colored
Band:	Brown leather band	Matching metal band	Black leather band
Price:	$ 349.-	$ 399.-	$ 299.-
Best option:	○	○	○



	Watch 1	Watch 2	Watch 3
Watch face:	Rectangular	Round	Rectangular
Case color:	Gold-colored	Gold-colored	Silver-colored
Band:	Brown leather band	Matching metal band	Black leather band
Price:	$ 349.-	$ 399.-	$ 299.-
Best option:	○	○	○

INCENTIVE-ALIGNMENT. We told the incentive-aligned respondents that 1 in 500 respondents would receive the smartwatch that he or she chose in a randomly-selected choice task as well as any difference between the displayed price and $500 (Ding 2007; Ding, Grewal, and Liechty 2005). Respondents who chose the outside option would receive $500 in cash. Incentive-aligned respondents watched a one-minute video describing the incentives (https://youtu.be/DBLPfRJo2Ho). To avoid confusing respondents, we used, for each experimental condition, a video that was consistent with that experimental condition. For example, if respondents were in the low-realism manipulation, the example choice sets shown in the video used lower-realism text-based images. In this way we measure the incremental impact of incentive alignment. Respondents who were not incentive-aligned were told that 1 in 500 would receive a cash bonus of $500. The $500 was not tied to their choices.

TRAINING VIDEO. Respondents assigned to the training video were required to watch a two-minute animated video describing the smartwatch category, the smartwatch attributes, and the choice tasks (https://youtu.be/oji_bw_oxTU). We matched the videos to the experimental cells. We chose not to force an equivalent two-minute delay for respondents who were not assigned to the training video, because such a delay does not represent common practice and might introduce a demand artifact. Our goal was to determine whether or not a training video is higher craft or lower craft. It might be higher craft because well-instructed respondents might understand the

task better; it might be lower craft if the additional instructions do not overcome the potential for respondent wear-out.

CETERIS PARIBUS INSTRUCTIONS. CBC formats assume that all product attributes that are not varied in the choice tasks are held constant in the choice tasks (Green and Srinivasan 1990). If respondents do not understand that such attributes are to be held constant, they might infer unobserved characteristics to be correlated with the attributes that are varied. For example, without ceteris paribus instructions, respondents might be more likely to infer that quality changes if prices change. We used the following instructions for respondents assigned to the (hypothesized) higher level of craft. Respondents in the (hypothesized) lower level of craft received no reminders to hold all other attributes constant.

> *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.*

THE EMPIRICAL EFFECT OF CRAFT ON ACCURACY AND PRECISION

The final sample of respondents who completed both waves of the study was 1,044 respondents split roughly equally among the experimental conditions in the orthogonal half replicate of the 2^4 design. (109 respondents always chose the outside option and were eliminated as not interested in smartwatches, at least not interested in the smartwatches in the conjoint design. The elimination of respondents was not correlated with any experimental manipulation.)

Roughly 500 respondents were assigned to each condition, e.g., realistic images vs. text-based images. Respondents took approximately 200 seconds to answer the choice tasks and this did not vary among experimental manipulations. The total survey took approximately 400-500 seconds to complete. Those respondents who were assigned to incentive alignment took approximately 69 seconds longer. Those respondents who were assigned to the training video took approximately 142 seconds longer. The longer duration corresponds to the length of the mandatory videos in these conditions. In approximately 25% of the choice tasks, respondents chose the no-choice outside option. This was slightly higher (3-5%) for more-realistic images, for those who saw the training video, and for those who were incentive aligned.

IMPACT OF CRAFT ON ACCURACY. Neither incentive alignment, the training video, nor ceteris paribus instructions affected the relative partworths substantially. For example, on average, differences in relative attribute importances between manipulations varied by but a few percentage points, usually by no more than one percentage point. On the other hand, realistic images mattered. Those respondents who were shown more realistic images were more likely to value differences in the watch band (40% vs. 27% relative importance of watch band) and less likely to value differences in color (17% vs. 22% relative importance of watch-face color). They

were also relatively less price sensitive (21% relative importance of price vs. 28% relative importance of price). Thus, it appears that only the realism of the images affects accuracy in our data.

Accuracy (the relative partworths) affects WTP, where WTP is interpreted as consumer surplus—the demand curve. Although, in McFadden's (2014) formulation, the relative partworths are the WTPs, there are issues with this interpretation when hierarchical Bayes or empirical Bayes analysis is used. First, the partworths are heterogeneous. This heterogeneity is inherent whether we sample from the hyper-distribution, from the full distribution of respondent-level partworths, or if we use the mean partworths for each respondent. Second, the conjoint study only collects data within certain price ranges. For example, our prices varied from $299 to $449. Samples from the hyper-distribution, samples within the full distribution, or even the mean partworths might imply WTPs outside this range. It would violate sound scientific principles to extrapolate beyond the price range that was used in the survey. Thus, we need a robust summary that does not violate these scientific principles. One method is to use robust statistics, such as medians. Another procedure, that is commonly applied, is to use a two-product simulator in which one product has the attribute level of interest and the other has the lower level of the attribute. The price that equalizes the predicted shares of both products can then be interpreted as WTP. We adopt this common practice for estimating the market's WTP.

Because only the realism of the images affects accuracy, only the realism of the images affects WTP substantially. As shown in Table 3, this impact can be dramatic suggesting that researchers should pay close attention to the realism of the images. The default of text-based formats may underestimate the willingness to pay for a product attribute.

Table 3. The Realism of Images in Conjoint Analysis Affects Estimated WTP

Calculated vis the Simulation Method	More Realistic Images	Text-Based Images
Round to Rectangular Face	$103	$39
Gold to silver color	$65	$59
Brown to black band	$130	$42
Metal to black band	$132	$4

IMPACT OF CRAFT ON ESTIMATION PRECISION ($\gamma^{estimation}$). The realism of the images and incentive alignment do not impact the precision that we estimate from the estimation data, $\gamma^{estimation}$. This is intuitive, $\gamma^{estimation}$ summarizes the internal consistency of the estimated logit model. When respondents are asked to choose among text-based descriptions, they might be extremely consistent in reporting their preferences for text-based descriptions. This does not necessarily

mean that the reported preferences for text-based stimuli describe how respondents would react in the marketplace.

Two aspects of craft affect estimation precision negatively. Estimation precision is significantly lower for respondents who saw the training video and those who were provided with the ceteris paribus instructions. The training video effect is intuitive. The extra time necessary to watch the training video may have been burdensome. It appears the time to watch the video was sufficiently burdensome so that respondents were less internally consistent in their choices in the estimation choice sets. The ceteris paribus instructions had a small, but statistically significant, negative impact on precision.

IMPACT OF CRAFT ON EXTERNAL VALIDITY PRECISION ($\gamma^{externally\ adjusted}$). When adjusted for external validity, precision was significantly better for more realistic images (vs. text-based images) and for incentive alignment (vs. no incentive alignment). For our data, these results imply that higher craft means using enhanced image realism and incentive alignment. Providing a training video or ceteris paribus instructions did not affect the adjustment based on external validation. However, overall the externally-adjusted precision in the training video condition is lower due to the lower estimation precision. Although significant, the effect is rather low in magnitude. Overall, there is no effect of ceteris paribus instructions; either they have little effect in general or respondents instinctively held all other attributes constant when making choices among the product profiles in our research setting.

The training video lowered the externally adjusted precision. For our data, it appears that the negative wear-out effect was larger than the (hypothesized positive) training effect. While our training video was professional quality, it was long relative to the total time of the questionnaire (142 incremental seconds out of approximately 500 seconds). We hypothesize that more concise training videos might have a more positive impact. Training videos might also be valuable for product categories that are harder for consumers to grasp. At minimum, our results caution conjoint-analysis analysts that training videos must be crafted and pretested carefully.

CRAFT AFFECTS MANAGERIAL RECOMMENDATIONS BECAUSE CRAFT AFFECTS PRECISION

We have already seen that craft in the realism of images affects accuracy and, by implication, estimated WTP. Craft also affects precision. We focus on craft in incentive alignment because craft in incentive alignment has a substantial effect on precision but not accuracy. By focusing on incentive alignment, we can illustrate a precision effect that is not confounded with an accuracy effect. (We return to realism in the images, and other aspects of craft, later in this paper. Image realism illustrates nicely the joint effect of precision and accuracy. No interactions among different aspects of craft were observed.)

IMPACT ON PREDICTED EQUILIBRIUM PRICES AND PROFITS. Table 4 compares estimated equilibrium prices and profits (shares*price). The results are based on partworths that are externally adjusted. The market is a two-product market in which the products vary on only the color of the watch

face. The estimated equilibrium prices and profits are for the products with the corresponding color.

In Table 4, improved craft predicts lower (and presumably more-correct) prices and profits. Lower craft (no incentive alignment) appears to overstate prices by about 3% and profits by about 5%. We obtain comparable effects for the shape of the smartwatch face and the type of smartwatch band. Externally adjusting the partworths tends to harmonize the results. These differences due to craft are substantially larger when relying on estimation precision only, which is addressed next.

Table 4. The Effect of Craft (Precision) on Equilibrium Prices and Profits

	Equilibrium Prices		Equilibrium Profits	
	Incentive Alignment	No Incentive Alignment	Incentive Alignment	No Incentive Alignment
Gold-colored Smartwatch	$311	$328	$99	$106
Silver-colored Smartwatch	$343	$347	$124	$129

EXTERNAL VS. INTERNAL PRECISION MATTERS WHEN CALCULATING PREDICTED EQUILIBRIUM PRICES AND PROFITS. Recall that craft in the realism of images had a two-fold effect; lower craft lowered both accuracy and precision. Table 3 illustrated that less-realistic images lowered predicted WTP for all smartwatch attribute-level differences—mostly because lower craft increased the estimated relative importance of price. For example, lower craft lowered the estimated WTP for gold vs. silver-colored smartwatches from $65 to $59 – about 9%. In contrast, a lower (externally-adjusted) precision suggests lower price sensitivity and higher equilibrium prices (e.g., see Table 2). The effects of accuracy (WTP) and precision (price sensitivity) might counteract one another.

We put together the joint effect of accuracy and precision and compare the differences in predicted equilibrium prices. For estimation precision, we expect the effect on accuracy to dominate. This is shown in Table 5a. Predicted equilibrium prices are roughly 25% lower for text-based images versus realistic images. Predicted equilibrium profits are 14% lower.

For externally-adjusted precision, text-based images lowered precision significantly, an effect that might counteract the effect of the lower accuracy (i.e., higher price sensitivity). This joint effect of accuracy and precision is shown in Table 5b. Predicted equilibrium prices and profits are, on average, higher for text-based images, but only by 6% and 11% respectively. (Note that predicted equilibrium prices and predicted equilibrium profits are higher in both conditions when based on externally adjusted precision vs. estimation precision. They increase by 47% and 26%, respectively.)

In our data, lower accuracy and lower precision counteract one another for craft based on the

realism of the images. But that will not always be the case. If lower craft were to lower the relative importance of price, then the two effects would reinforce one another. Lower craft would have an even bigger effect on managerial recommendations and on patent or copyright valuations. The key message in Table 5 is that researchers should pay attention to craft and to external-validity adjustments to precision.

Table 5. Adjustments to Reflect External Validity Matter Managerially (Example Where Decreased Accuracy and Decreased Precision Counteract).

Table 5a	Equilibrium Prices		Equilibrium Profits	
Results for Estimation Precision	More Realistic Images	Text-Based Images	More Realistic Images	Text-Based Images
Round Smartwatch Face	$233	$194	$79	$78
Rectangular Smartwatch Face	$317	$219	$124	$96

Table 5b	Equilibrium Prices		Equilibrium Profits	
Results for Externally-Adjusted Precision	More Realistic Images	Text-Based Images	More Realistic Images	Text-Based Images
Round Smartwatch Face	$298	$350	$92	$117
Rectangular Smartwatch Face	$386	$377	$134	$134

IMPACT ON THE STRATEGIC RECOMMENDATIONS FOR THE ATTRIBUTE LEVELS OF A PRODUCT. The description of the equilibrium in attributes and prices is beyond the scope of this paper. For more details on the theory, see Hauser, Eggers, and Selove (2016). They show for the magnitudes of precision that we find in our empirical data, the innovator would be advised to launch the more preferred silver-colored smartwatch. The follower's actions depend upon the precision that the follower estimates for the market. In particular, a follower would be advised to launch a:

- Gold-colored smartwatch if the true precision were higher
- Silver-colored smartwatch if the true precision were lower

We get similar effects for the watch face and the watch band.

INCREASED SAMPLE SIZE DOES NOT COMPENSATE FOR LOWER CRAFT

It is tempting to equate precision, as defined in this paper, with precision of the partworth estimates. This is an incorrect interpretation. Precision (scale of the partworths) is a different concept than the standard errors of the estimates of the partworths, or the related concepts in Bayesian analysis.

To illustrate that they are different concepts, we re-estimated all of our models using a randomly selected 50% sample. On average, the standard errors of the estimates of $\gamma^{estimation}$ and $\gamma^{externally\ adjusted}$ were 32% lower for the full sample compared to the random half sample, but the magnitudes of the estimates were comparable. Averaged over all conditions, the estimation precisions were within less than 1%, and the externally-adjusted precisions within 3%, when we compare estimates based on the full sample to estimates based on a random half of the sample.

Sample size does not overcome lower craft!

DISCUSSION AND SUMMARY

THEORY. Accuracy matters, but its effect has been well-studied. Accurate relative partworths are important for deciding which attribute levels to include in products and for calculating willingness-to-pay (WTP) as input to other managerial recommendations (and in litigation, as input to damages experts). However, precision (scale) matters as well. Precision affects directly predictions of equilibrium prices and profits and, as a result, recommendations on the strategic selection of attribute levels for products. The market reacts according to true precision (γ^{true}), but analysts make recommendations based on estimated precision. Recommendations vary dramatically depending upon whether the recommendations are based on estimation precision ($\gamma^{estimation}$) or externally-adjusted precision ($\gamma^{externally\ adjusted}$).

CRAFT. Craft affects both accuracy and precision. For the situation we examined empirically, we found:

- More realistic images increased both accuracy and precision.
- Incentive alignment increased precision, but had little effect on accuracy.
- Training videos had no effect on accuracy, but appear to decrease both estimation precision and externally-adjusted precision.
    - In our data, the training videos were time-consuming for consumers to watch and may have led to respondent wear-out. The wear-out effect might have been stronger than the training effect.
    - Well-designed training videos might enhance precision. One must craft and pretest such videos carefully.
- Ceteris paribus instructions had no substantial effect on either accuracy or precision. This result might be due to the fact that we used instructions that mimicked the status quo on the market, which consumers might have assumed implicitly even without instructions.

- Increased sample size does not compensate for lower craft.

MANAGERIAL RECOMMENDATIONS. Craft affects managerial recommendations. For the situations we examined empirically, we found:

- The realism of the images changed the relative importance of the attributes and decreased the relative importance of price. WTPs were higher for more realistic images.
- Predicted equilibrium prices and profits depend upon craft and whether or not the partworths are adjusted for external validity.
- Managerial recommendations may change if they are made using precisions as estimated from the conjoint analysis data or if they are made using precisions adjusted for external validity.
    - In some cases, the effect of craft on accuracy and precision counteract one another. In other cases, the effects of craft on accuracy and precision reinforce one another.
    - Because the directional impact cannot be predicted ahead of time, higher craft is advised.
- The correct strategic selection of attribute levels for products depends upon accuracy and precision, and, hence, craft.

SUMMARY. Craft matters! High craft avoids making the wrong (or costly) managerial recommendations. High craft avoids estimating incorrect valuations for copyrights and patents.



REFERENCES

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

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.

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 McFadden (2013), "The Role of Conjoint Surveys in Reasonable Royalty Cases," *Law360*, October 16.

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.

Eggers, Felix, John R. Hauser, and Matthew Selove (2016), "Precision Matters: How Craft in Conjoint Analysis Affects Price and Positioning Strategies,"(Cambridge, MA: MIT Sloan School of Management).

Green, Paul E., and V. Srinivasan (1990), "Conjoint Analysis In Marketing: New Developments With Implications For Research and Practice." *Journal of Marketing*, 54, 4, (October), 3-19.

Hauser, John R., Felix Eggers, and Matthew Selove (2016), "The Strategic Implications of Precision in Conjoint Analysis," (Cambridge, MA: MIT Sloan School of Management).

McArdle, Meghan (2000), "Internet-based Rapid Customer Feedback for Design Feature Tradeoff Analysis," S. M. Thesis. (Cambridge MA: MIT Sloan School of Management).

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.

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

Orme, Bryan K. (2010), *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research, 2E.* (Research Publishers LLC: Madison WI).

Sawtooth Software (2009). http://www.sawtoothsoftware.com/support/technical-papers/hierarchical-bayes-estimation/cbc-hb-technical-paper-2009.

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

Toubia, Olivier, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, 22, 3, (Summer), 273-303.

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.