WEB APPENDIX A: VERTICALLY-DIFFERENTIATED PRODUCT CATEGORIES (PEARSON CORRELATION BETWEEN AVERAGE USER RATINGS FROM AMAZON.COM AND *CONSUMER REPORTS* QUALITY SCORES)

Baby Monitors: Audio Models (0.58)
Bike Helmets: Youth (0.06)
Blood Glucose Meters (-0.99)
Blood Pressure Monitors: Arm Models (0.17)
Blood Pressure Monitors: Wrist Models (-0.27)
Camcorders (0.41)
Carbon Monoxide-Smoke Alarms: Interconnected Models (0.57)
Carbon Monoxide-Smoke Alarms: Stand-Alone Models (0.47)
Car-seats: 40 lb. Harness Capacity (0.35)
Car-wax: Liquid or Gel Wax (0.82)
Car-wax: Paste wax (0.76)
Clothes Dryers (0.61)
Coffee Makers: 10- to 12-Cup Brew and Dispense Models (-0.29)
Coffee Makers: 8 to 10 Cup Grind and Brew Models (0.42)
Coffee Makers: 8 to 14 Cup Drip Models with Carafe (0.23)
Computer Monitors: 22 to 26 inch Models (-0.36)
Computer Tablets: 7 to 8 inch Models (0.01)
Computer Tablets: 9 to 12 inch Models (-0.92)
Computer: 11 inch Models (0.30)
Computer: 13 inch Models (-0.02)
Computer: 14 inch Models (0.81)
Computer: 15 to 16 inch Models (0.09)
Computer: 17 to 18 inch Models (0.13)
Conventional Dishwasher (0.09)
Cordless Impact Drill/Drivers (0.41)
Cordless Phones (0.65)
Cordless Phones with Answerer (-0.77)
Cordless Screwdrivers (0.03)
Cribs: Stationary Sides (0.55)
Dehumidifiers: Large Capacity (0.31)
Dehumidifiers: Medium Capacity (-0.59)
Digital Cameras: Compact Models (0.99)
Digital Cameras: Subcompact Models (0.75)
Digital Cameras: Super Zoom Models (-0.20)
Digital Picture Frames: 6 to 8 inch Display (0.09)
Digital Scales (-0.43)
Dishwater Detergents (-0.42)
E-book Readers: 6 to 7 inch Display (0.06)
Electric Blenders (0.95)
Electric Mixers (0.42)
Electric Razors: Foil Type (0.63)
Electric Razors: Rotary Type (0.08)
Elliptical Exercisers (-0.44)
Food Processor (0.45)
Full Size Carpet Cleaning Machines (0.65)
Garbage Disposers: Continuous Feed (0.09)
Gas Grills: Medium Size Models (0.50)
Gas Grills: Portable or Small Models (-0.22)
General Use Cordless Drill/Drivers (-0.25)
GPS: 3.5 inch screen size (0.04)
GPS: 4.3 inch screen size (0.40)
GPS: 4.7 to 7 inch screen size (-0.57)
Hair Dryers (-0.03)
Hair Dyes (0.09)
Headphones: Home/Studio Style (0.16)
Headphones: Portable (0.40)
Heart Rate Monitors (0.25)
High Chairs (0.37)
Home Theatre Systems: Sound Bars (0.49)
Home Theatre Systems: With Blu-ray Player (0.55)
Home Theatre Systems: Without Video Player (0.60)
Humidifiers: Tabletop (0.17)
Juicers (0.42)
Kitchen Cookware: Non-stick (-0.43)
Kitchen Cookware: Uncoated (0.28)
Kitchen Knives: Fine-edge blades (0.53)
Kitchen Ranges: 30 inch Smooth Top (0.85)
Laundry Detergent: Conventional (0.59)
Laundry Detergent: High Efficiency (0.40)
Leaf Blowers: Electric Handheld (0.60)
Leaf Blowers: Gasoline Handheld (0.40)

Light Bulbs: 65 Watt Equivalent R30 Flood/Reflector (0.47)
Light Use Cordless Use Drill/Drivers (0.40)
Microwave Ovens: Large (-0.91)
Microwave Ovens: Midsized (-0.78)
Mono Bluetooth Headsets (-0.33)
MP3 Players: Media Players (-.0.04)
MP3 Players: Music Players (-0.08)
Paper Shredders: Pull Out Console (0.16)
Paper Shredders: Wastepaper Basket (0.93)
Paper Towels (0.08)
Portable Air Purifiers (0.04)
Printers: Black and White Laser (0.39)
Printers: Inkjet (0.80)
Refrigerators: French-door bottom-freezer (0.90)
Safety Gates: Hardware Mounted Gates (0.79)
Security Software: Pay Security Suites (0.52)
SLR Lenses: Standard Zoom (-0.08)
SLR Lenses: Super Zoom (-0.38)
SLR Lenses: Tele Zoom (0.82)
Snow Blowers: Single-stage Electric Models (0.43)
Space Heater: Electric (0.58)
Space Heater: Electric Fan Force (0.52)
Sprint Nextel Cell Phones (-0.66)
Steam Irons: Conventional (-0.39)
Steam Irons: Steam Ironing Systems (-0.43)
Steam Mops (0.01)
String Trimmers: Battery (-0.19)
String Trimmers: Electric (0.31)
String Trimmers: Gasoline (-0.35)
Strollers: Double side-by-side (0.37)
Strollers: Single-combo (-0.16)
Strollers: Single-traditional (0.62)
Strollers: Single-umbrella Style (0.26)
Sunscreen: SPF 30 (0.47)
Thermometers: Infrared (-0.55)
Thermostats: Seven-Day Models (-0.50)
Tire Pressure Gauges (0.33)
Toasters: Toaster Ovens/Broilers (-0.73)
Tougher Job Cordless Use Drill/Drivers (0.30)
Treadmills: Folding (-0.26)
Treadmills: Folding: Budget (-0.18)
Treadmills: Non-Folding (0.81)
TV's: LCD TV Models (1.00)
TV's: Plasma TV Models (-0.19)
Vacuum Cleaners: Bagged (0.74)
Vacuum Cleaners: Bagless (0.02)
Water Filters: Carafes (-0.52)
Water Filters: Faucet-mounted Filters (0.24)
Water Filters: Under Sink Filters (-0.40)

WEB APPENDIX B: ATTRIBUTE WEIGHTING ANALYSIS

We were able to realize a full set of attribute scores for 1,059 products in 111 categories, or 83% of products in 93% of categories. Most attribute scores are on a 1-5 scale where higher numbers are better. For a small number of attributes (e.g., yearly operating cost), higher numbers are worse, so we reverse coded these attributes. Next, we z-scored all attribute scores within category to make all scales comparable. We calculated covariance matrices between all attributes and computed the percentage of positive and negative covariances in each product category. Similar to Curry and Faulds (1986), covariances were primarily positive (72% of covariances, averaged across categories), suggesting that differences in the weighting scheme do not explain the low correlation between average user ratings and composite *Consumer Reports* scores.

To verify and quantify this result, we ran a Monte Carlo simulation. The simulation can be thought of as modeling a "virtual consumer" that rates each product in a category based on random set of weights sampled from a uniform distribution. The simulation assumes that the virtual consumer has the same attribute scores as *Consumer Reports* but idiosyncratic weights, and the consumer's score for each product is the weighted sum over attributes. We then calculated the correlation between the consumer's product scores based on the weighted sum of attribute scores, and the overall *Consumer Reports* score for that product. The process was repeated for all categories. We ran 10,000 simulations, or 10,000 virtual consumers, then analyzed the data by category and by virtual consumer.

Averaged across the 10,000 virtual consumers, the median category correlation was 0.83 and the mean correlation was 0.75. The distribution over categories was highly left skewed with the majority of categories (87%) having correlations above 0.60. Looking across categories, the average virtual consumer's median correlation was 0.83 and mean correlation was 0.75. The

virtual customer with the lowest correlation to *Consumer Reports* scores had a median correlation of 0.78 and a mean correlation of 0.69. The virtual customer with the highest correlation to *Consumer Reports* scores had a median correlation of 0.88 and a mean correlation of 0.81.

Thus, the plausible range of values for the correlation between user ratings and *Consumer Reports* scores, across categories, assuming consumers have different weights than *Consumer Reports*, but score the attributes the same, is 0.7-0.9. Note that this is a highly conservative test because it assumes that consumer weights are independent of *Consumer Reports*' weights, which is unlikely. This result shows that variation in the weighting scheme when aggregating attributes does not explain the low correlation between user ratings and *Consumer Reports* scores.

WEB APPENDIX C: CONSUMER STUDIES

Study 1

Participants and Procedure. We recruited 152 U.S. residents from Amazon Mechanical Turk and paid them $0.65. Each participant responded to five vertically-differentiated product categories randomly sampled from the 10 product categories with the most items in our database of market data (air purifiers, digital cameras, coffee makers, food processors, steam irons, kitchen knives, GPS navigators, cordless telephones, printers, and strollers), and presented in random order. For each category, participants read the following instructions: "Imagine you were looking to buy a [category name]. Many consumers may go to a website such as Amazon.com to learn from online ratings and reviews provided by other consumers who have previously purchased and used the product. Imagine you are looking at [category name], what do you hope

to learn from these online ratings and reviews? Please list five specific things." Participants entered their responses into free-response text boxes. Next, we gave them a list of the dimensions rated by *Consumer Reports* in that category and they were told, "here are some additional things we came up with that you might want to learn about [category name] from online consumer ratings and reviews. Please indicate the information in your list that is unique, that is, indicate the information in your list that is not mentioned in ours." Finally, we gave participants an integrated list of the dimensions covered by *Consumer Reports* and any unique dimensions they had listed, and we asked them to rank the list from most to least important. Two hypothesis-blind research assistants individually coded the open-ended responses. Coders were instructed to first determine whether the dimensions listed by respondents were covered by *Consumer Reports*. Coders made these decisions based on a document containing a complete list of dimensions evaluated by *Consumer Reports* in each category as well as the description for each dimension taken from the *Consumer Reports* website. If the coders concluded that the dimension was not covered by *Consumer Reports*, they assigned it to one of nine other categories. The coding instructions given to coders are shown immediately below.

Coding Instructions.

Specific Quality Dimensions Covered by *Consumer Reports*:

- Responses in this category are ones that are covered by the given list provided by *Consumer Reports* for the given product. (Sometimes it is not totally evident that a response is equivalent to an attribute rated by *Consumer Reports*, although it is in actuality. For example, for air purifiers, if a response was "Filter Micron Size," that would be the same as "Dust" if you were to read the *Consumer Reports* description of the "dust" attribute. Therefore, please read the *Consumer Reports* description for each of the

attributes rated by *Consumer Reports* before coding each category.)

Specific Quality Dimensions Not Covered by *Consumer Reports*:

- Responses include technical abilities of the products that are not covered by *Consumer Reports*. Common concerns are compatibility with other devices and overall energy use of the product.

Quality (Generic):

- These responses are those which cover the general, overall quality of the product, or the general quality of materials used to make the product and the reliability of the item as it serves the customer. Note that it refers to "general responses of quality," not quality that refers to any particular attribute.

Price/Value:

- Responses in this category refer to things such as the overall best price or value for a given product compared to other competitive products. This code also includes the costs associated with the product compared to others on the market.

Subjective Evaluation:

- This category covers concerns about color options, style, and product appeal. Personal product satisfaction as well as aesthetic appeals are also covered in this category.

Durability/Longevity:

- This category includes concerns about the product's ability to resist every day wear, tear, decay, and use. It also includes the duration that it can withstand these effects, that is, its estimated useful life.

Brand:

- This category covers responses that explicitly mention looking for information about the brand.

Warranty/customer support:

- This code covers if there are any warranty options for the products as well as concerns about product support.

Ease of Use Not Covered by *Consumer Reports*:

- This code includes reports about the ease of programming, operating and transporting the given product if it is not covered by *Consumer Reports* for a given product category.

Uncodable:

- These include responses that are not covered by *Consumer Reports* and do not fit in any of the given categories, or things that we do not understand.

The coders agreed in 85% of cases (Cohen's Kappa = .81). In cases where they disagreed, they met and agreed upon a single code. The rectified coding was used for all subsequent analyses.

Results. As can be seen from figure 3 in the paper, the most important dimension to consumers is covered by *Consumer Reports* 71% of the time, and all dimensions mentioned by consumers are covered by *Consumer Reports* 55% of the time. Thus, objective dimensions covered by *Consumer Reports* were by far the most common reason for respondents to consult user ratings and reviews on Amazon.com. Another common reason was to learn about the price or value of a product. Some consumers reported consulting user ratings and reviews to learn about more subjective evaluations, but much less frequently (3.6% of the most important dimensions and 6.3% of all dimensions). These results suggest that consumers consult user ratings for vertically-differentiated product categories primarily to learn about technical dimensions of quality that are amenable to objective tests and are covered by *Consumer Reports*.

Study 2

The goal of study 2 was to quantify consumers' reliance on the average user rating as a cue for quality. We asked consumers to search for pairs of products on Amazon.com and then to rate which product they thought *Consumer Reports* would rate higher. Because these products vary naturally in terms of average user ratings and other cues (e.g., price and number of ratings), we were able to test how variation in the average rating influences quality judgments, and compare consumers' reliance on the average user rating to other available quality cues. To avoid any demand effects we designed the search and rating task to be as realistic as possible and we gave participants no training and minimal instructions.

Participants and Procedure. We recruited 304 U.S. residents from Amazon Mechanical Turk and paid them $0.65 for completing the study. We first informed participants that they would be provided with pairs of products within a category along with the URLs for the products at Amazon.com. We asked participants to inspect the product web pages and judge which product they thought *Consumer Reports* would rank higher. To ensure that participants understood that they were evaluating quality as measured by *Consumer Reports*, we provided the following description on the instruction page: "Expert ratings like those generated by *Consumer Reports* magazine are generated by engineers and technicians with years and sometimes decades of expertise in their field. They live with the products for several weeks, putting them through a battery of objective tests using scientific measurements, along with subjective tests that replicate the user experience. All models within a category go through exactly the same tests, side by side, so they're judged on a level playing field, and test results can be compared." After reading the instructions, participants saw the first pair of products. They were asked to click on the links and

examine the products. The links took the participants to the actual live web pages for the products on Amazon.com so that they saw all information on the products as any shopping consumer would. Participants were then asked to rate "how you feel experts at *Consumer Reports* magazine would likely rate the quality of the products relative to each other." The rating scale ranged from 1 (Product A would be rated as higher quality) to 10 (Product B would be rated as higher quality). After completing the ratings, participants were provided with the link to the next set of products. Each participant completed the quality rating task for one pair of products in each of eight product categories. The two products were sampled at random from a pre-determined set as specified below, and category presentation order was randomized as well.

Product Selection. We initially selected the 10 product categories with the most items from the database of marketplace data used above. We then excluded MP3 players and kitchen knives because many items in these categories were not comparable to one another. For example, in the kitchen knife category, some products referred to single knives while others referred to knife sets. We then selected one product from each brand represented in the category. The final set of stimuli resulting from this procedure included four printers, seven digital cameras, two GPS navigators, four cordless phones, 16 coffee makers, 10 steam irons, eight food processors, and seven strollers.

Results. For each product pair, we computed the difference between product A and product B in average user rating, number of user ratings, and price. We collected this data from the Amazon.com website right before launching the study. It is important to note that while we collected these three variables from each of the respective product web pages prior to the study for use as predictor variables, participants were exposed to the full array of information on the product web pages, thereby enhancing external validity. To measure the extent to which sample

sizes are sufficient to discriminate average ratings of two products in a pair, we computed the Satterthwaite approximation for the pooled standard error (hereafter referred to as "pooled standard error"), which is a function of the sample sizes and the variation in user ratings of products A and B ($SE_{Pooled} = \sqrt{[(VAR_A/N_A) + (VAR_B/N_B)]}$). As pooled standard error decreases, sample size becomes more sufficient relative to variability. To make effect sizes comparable across cues we standardized all variables by product category such that they had a mean of zero and a standard deviation of one. We regressed consumers' judgments of quality on (1) the difference in average user ratings, (2) the pooled standard error of the difference in average user ratings, (3) the interaction between the difference in average user ratings and the pooled standard error of the difference in average user ratings, (4) the difference in the number of user ratings, and (5) the difference in prices. Results are summarized in table 1 in the paper (see consumer study 2). Quality judgments were more strongly related to differences in average user rating (b = 0.34, CI_{95}: [0.31, 0.38]) than to differences in price (b = 0.21, CI_{95}: [0.17, 0.24]) and differences in the number of ratings (b = 0.14, CI_{95}: [0.10, 0.18]). The 95% confidence interval for the average user rating did not overlap with the 95% confidence intervals for price and the number of user ratings, indicating that the average user rating explained significantly more variation in judgments of quality. The effect of pooled standard error (b = 0.00, CI_{95}: [-0.04, 0.04]) and the interaction effect between the difference in average user ratings and pooled standard error (b = 0.03, CI_{95}: [-0.01, 0.07]) were not statistically significant. Thus there is no evidence that consumers took the sufficiency of sample size into account when making quality inferences.

Study 3

This study was a replication of the second consumer study, with three changes. In the previous study, we asked participants to evaluate "quality as measured by *Consumer Reports*". However, when consumers predict quality in a real shopping context, they may use a different lay definition of quality and this may affect cue utilization. Thus, in the current study we asked participants to infer quality without any additional elaboration. Second, before making the quality judgment, we asked consumers to copy the price, the average rating, and the number of ratings into a table. Given the lack of a moderating effect of pooled standard error on quality inferences in the previous study, we wanted to draw participants' attention to the distribution of ratings (by asking them to copy the number of reviews) to see if this increased sensitivity to pooled standard error. Finally, in addition to a quality judgment, we asked participants to make a hypothetical purchase decision.

Methods. One-hundred-forty-five respondents from Amazon Mechanical Turk were paid $0.65 to complete the study. The procedures and instructions were the same as in study 2, except for the following changes. First, instead of asking participants to make an inference about *Consumer Reports* scores, we asked consumers to make an inference about quality without any further elaboration. Second, before making the quality judgment participants were asked to copy the average rating, number of ratings, and price for both products into a table. This table was displayed on the screen as participants made the quality judgment. Third, after making the quality judgment, participants were asked to choose between the products on a scale from 1 (I would definitely choose product A) to 10 (I would definitely choose product B).

Results. We examined quality judgments and purchase likelihoods with the same regression model as in the previous study. Again, as shown in table 1 in the paper, quality judgments were more strongly related to differences in average user rating (b = 0.40, CI_{95}: [0.35,

0.45]) than to differences in price (b = 0.13, CI_{95}: [0.08, 0.17]) and differences in the number of user ratings (b = 0.22, CI_{95}: [0.17, 0.26]). The reliance on average rating relative to price was stronger in this study compared to the previous study. The regression coefficient for the average user rating was over three times that of price, compared to just under two times in the previous study. Also the reliance on the number of user ratings relative to price was stronger in this study compared to the previous study. Whereas consumers relied more on price in the previous study, they relied more on the number of user ratings in this study. This is likely because participants in the previous study were not forced to copy the predictive cues into a table. On this account, consumers do not naturally attend to the number of ratings as much as price when evaluating products online. Although we attempted to draw participants' attention to the distribution of ratings by having them copy the number of ratings into a table, participants' reliance on the average user rating was again not moderated by pooled standard error (b = 0.00, CI_{95}: [-0.05, 0.06]). The effect of pooled standard error was also not significant (b = 0.04, CI_{95}: [-0.01, 0.09]).

For purchase likelihood, there was a significant positive effect of average user rating (b = 0.35, CI_{95}: [0.30, 0.40]), a significant negative effect of price (b = -0.41, CI_{95}: [-0.46, -0.37]), and a significant positive effect of the number of user ratings (b = 0.24, CI_{95}: [0.19, 0.28]). Again, there was no effect of pooled standard error (b = 0.02, CI_{95}: [-0.03, 0.07]), nor was there an interaction effect of average rating with pooled standard error (b = -0.01, CI_{95}: [-0.07, 0.05]).

To test for mediation, we included judged quality as a predictor in the model. The effect of perceived quality on purchase likelihood was positive (b = 0.56, CI_{95}: [0.51, 0.61]), such that products perceived to be of higher quality were more likely to be chosen than products perceived to be of lower quality. The effects of average user rating (b = 0.13, CI_{95}: [0.09, 0.18]) and number of user ratings (b = 0.10, CI_{95}: [0.06, 0.14]) were again positive but weaker than in the

model excluding perceived quality. This suggests that the average user rating and number of user ratings have an influence on expected quality, which in turn affects purchase likelihood. The effect of price was again negative and stronger than in the model excluding perceived quality (b = -0.46, CI_{95}: [-0.50, -0.42]). This pattern of results is consistent with the dual role of price. Price feeds into purchase decisions as a "negative indicator of value" but also as a "positive indicator of quality." Thus, after controlling for perceived quality, the effect of price on purchase likelihood should be more negative, as it is. To examine whether the indirect effects of the average user rating, the number of user ratings, and price on purchase likelihood through perceived quality are significant, we used the bootstrapping procedure of Preacher and Hayes (2004; see Zhao, Lynch, and Chen 2010). The indirect effects were significant for the average user rating (b = 0.21, CI_{95}: [0.14, 0.29]), the number of user ratings (b = 0.15, CI_{95}: [0.12, 0.18]), and price (b = 0.12, CI_{95}: [0.08, 0.16]), indicating significant mediation through perceived quality.

Study 4

In the previous studies, we maximized realism by having consumers visit the Amazon.com website and evaluate products in a naturalistic way. While heightening realism, these designs suffer from extraneous sources of variation that could challenge our interpretation of the results. We therefore conducted an additional study where we manipulated the average user rating, the number of user ratings, and price, while keeping the stimuli otherwise identical.

Methods. Three-hundred-fourteen U.S. residents were recruited from Amazon Mechanical Turk and were paid $0.50 to complete the study. Participants rated the relative quality of two brands in eight product categories. We used the same product categories as in

consumer studies 2 and 3. Again, categories were presented in randomized order for each participant. We did not inform participants about brand names but simply referred to the products as "Brand A" and "Brand B." For each product, we presented participants with a table containing information about price, the average user rating, and the number of user ratings. Participants predicted how *Consumer Reports* experts would rate the relative quality of both products, using the same instructions and rating scale as in consumer study 2.

The brands differed along three dimensions: the average rating, price, and the number of ratings. Brand A could score either low or high on each of the three dimensions and the levels for Brand B were chosen to differ from those of Brand A, yielding eight possible brand pairs in each category. For instance, if Brand A had a low average user rating, Brand B had a high average user rating. The low and high levels for each product were chosen based on the database of products used in the analyses of market data. Low levels corresponded to the 25^{th} percentile in the category and high levels to the 75^{th} percentile in the category. For example, for digital cameras the low and high levels for price were $104.48 and $212.00, the low and high levels for the average user rating were 3.55 and 4.10, and the low and high levels for the number of ratings were 24 and 212. Each participant was assigned to eight comparisons, one in each category. Assignment to comparisons was randomized for each participant.

Results. For each dimension in each product category, we coded whether the level of Brand A on the dimension was low (-1) or high (1). We then estimated a regression model in which we entered Brand A's average rating level, price level, and number of ratings level as predictors. We also entered the interaction between average rating level and number of ratings level. Since we did not provide participants with the full distribution of ratings to allow for the calculation of pooled standard error, we used the interaction of the average user rating and the

number of ratings as a proxy for sensitivity to pooled standard error. As shown in table 1 in the paper and replicating the previous studies, quality judgments were more strongly related to the average user rating (b = 0.67, CI_{95}: [0.64, 0.70]) than to price (b = 0.02, CI_{95}: [-0.01, 0.04]) and to number of ratings (b = 0.22, CI_{95}: [0.19, 0.25]). In fact, the effect of price was not significant in this study. Also replicating previous studies, the interaction between the average user rating and number of ratings (a proxy for the pooled standard errors of the difference in average user ratings) was not significant (b = 0.01, CI_{95}: [-0.02, 0.04]).

REFERENCES

Preacher, Kristopher J., and Andrew F. Hayes (2004), "SPSS and SAS Procedures for Estimating Indirect Effects in Simple Mediation Models," *Behavior Research Methods, Instruments, & Computers*, 36 (4), 717-31.

Zhao, Xinshu, John G. Lynch, and Qimei Chen (2010), "Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis," *Journal of Consumer Research*, 37 (2), 197-206.