Journal of Management Science and Business Intelligence, 2018, 3–1
July 2018, pages 15-30
doi: 10.5281/zenodo.1246313
http://www.ibii-us.org/Journals/JMSBI/
ISBN 2472-9264 (Online), 2472-9256 (Print)

Quantitatively Exploring Human Preference

Matt Vance[1, *], John Carstens[1], Michael Gasper[1], Phillip Parker[1]

[1]Oklahoma State University, Department of Aviation and Space, 319 Willard Stillwater, OK 740078

*Email: matt.vance@okstate.edu

Received on February 10, 2018; revised on May 12, 2018; published on May 13, 2018

Abstract

Qualitative preference is frequently used to make significant, programmatic choices between competing suppliers and products. Our research helped identify the degree of quantitative granularity, which must be present between criteria for defensible choices, by exploring how small of a difference humans are able to reliably perceive. The response options, the type of response option (either physical weight differentiation or web-based shape/size differentiation), and the granularity of these responses were the quantitative focus of our research. From this quantitative basis, we were able to postulate corresponding qualitative-preference implications, impacts, and levels of reliability in the responses. Two primary data collection methods were designed to contrast each other: an in-person convenience sample taken on the Oklahoma State University (OSU) campus with physical weights, and an online survey of randomly sampled respondents based on pair-wise, visual-size differences of circle areas. Weber's Law was the primary analytical exploratory tool employed. Our research showed (a) the larger a given pair-wise comparison is in magnitude (weight or size), the more reliably human preference for the larger (either heavier or larger in size) commodity can be measured, (b) visual perception of area difference was consistently more accurate than weight difference – even with small differences of just 10%, and (c) if 95% reliability in choice between alternatives was desired then the perceived difference needed to be a factor of 3.0 (the larger choice needed to be either 200% heavier or 200% larger in area).

Keywords: Preference, Weber's Law, Quality Function Deployment, Choice

1 Introduction

Qualitative preference is frequently used to make significant programmatic choices between competing suppliers and products. Our research helped identify the degree of quantitative granularity that must be present between criteria for reliable, defensible choices. The macro purpose of this research was to explore how small of a difference humans are able to reliably perceive. The impetus for our research was the principal investigator's corporate association and use of a popular, late-1980's, Total Quality Management (TQM) tool – Quality Function Deployment (QFD).

QFD is a both a qualitative and quantitative tool. It allows the analyst to quantify what would typically be subjective judgement. The qualitative preference scale QFD uses is based on a quantitative factor of three: "None" = 0, "Weak" = 1, "Moderate" =3, and "Strong"= 9. The times-three factor is the specific target of this research – Does a human require a tripling in magnitude to reliably distinguish preference?

The QFD 0-1-3-9 preference scale discussed in Fiorenzo and Alessandro (1999) is the root scale that started this research discussion; however, it is not the only scale or difference set explored. A preference factor of three seemed obviously distinguishable and is significantly larger than the literature suggests as a "just noticeable difference" (Britt & Nelson, 1976). The type of response options (either physical weight or web-based size differentiation) and the granularity of these response options were the quantitative focus of this research. From this quantitative basis, we were able to postulate corresponding qualitative preference implications and impacts.

In the mid-1980s, QFD was popularized in the United States (U.S.) auto industry by the Ford Motor Company; it then migrated into aerospace shortly thereafter. QFD's primary attraction was capturing customer requirements in the embryonic stages of a development program. Later in the program development life-cycle, QFD Concept Selection proved to be a useful method in choosing among competing alternatives. Throughout the 1990s, both of these methods were successfully utilized at McDonnell Douglas and then the Boeing Corporation by the principal investigator and teams of trained colleagues in a large variety of internal and external, fee-for-service applications to explore the foundations of small, prototypical programs and substantial, multimillion-dollar, new-start programs.

A flagship QFD application, which the principal investigator led in 1995, explored the feasibility of incorporating an "Advanced Common Flight Deck" into the then MD-80, MD-95, MD-10 and MD-11 jetliners (Vance, 1995). The application of QFD as a decision tool flowed as a process. First, the overall design criteria and their valuation/weighting were agreed upon by the Douglas Aircraft design team. Second, the design team was partitioned into four, parallel teams of subject matter experts who were assembled to examine segregated components of the flight

deck. The four, sub-teams were responsible for ergonomics, instrument panel, overhead panel, and pedestal. Third, these sub-teams generated concepts from their respective flight deck component areas to meet the joint criteria; each sub-concept was evaluated and scored against the joint criteria. Fourth, the sub-teams reassembled to propose five, overall "Advanced Common Flight Deck" concepts created from the sub-concepts. Fifth, the overall concepts were then evaluated against performance, cost, and schedule criteria. The key in this evaluation was the ability to both forward-fit a concept into new products and the ability to retro-fit a concept into the current product line. Sixth, the results of the methodical, step-by-step, qualitative QFD analysis (with quantitative results) were presented to executive management. Evaluating the ability, opportunities, and risks to introduce advanced, flight deck commonality across the Douglas Aircraft product line was influential in the McDonnell Douglas executive decision to not invest further in Douglas Aircraft.

Figure 1 shows four, numerical-preference scales mapped against the subjective "None", "Weak", "Moderate", and "Strong" preference scale.

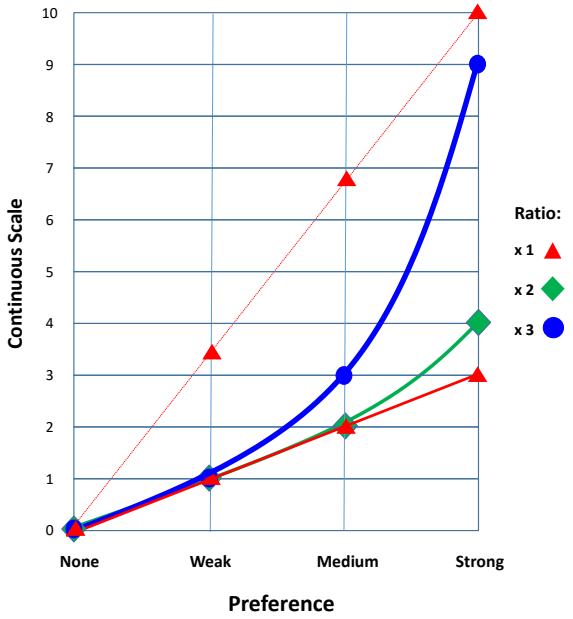

Fig 1. Preference Scales. The purpose of this graph is to illustrate two linear and two nonlinear numerical, quantitative scales associated with the QFD qualitative "None", "Weak", "Medium", "Strong" preference scale.

Figure 1 shows categorical preference along the x-axis equated with a continuous numerical scale along the y-axis. It is important to remember that a continuous scale can be infinitely divided if needed and therefore supports any desired mathematical operation. Continuous scales facilitate ratio data which means that the average and, if desired, the standard deviation of the preference scale associated scores can reliably be computed and mathematically compared.

The red lines show two variations of a linear preference scale. A linear preference scale requires the user to agree each increase in preference is of equal magnitude; in other words, moving from "None" to "Weak" is equivalent to moving from "Weak" to "Medium", or moving from "Medium" to "Strong". As the preference increases on the lower red scale, each increase is associated with a 1.0 numerical increase. The upper (dotted line) red scale shows for each preference increase the numerical increase is 3⅓. While the scales have a different numerical association, the only difference in their application would be the magnitude of the resulting scores. Both are ratio scales with a factor of 1.0 which means each step increase is an additional 1.0 times the previous level.

As shown in Hoyle, Chen, Wang & Gomez-Levi (2010), preference is almost always not linear and some form of nonlinearity is more representative of human choice which means a "Strong" preference should be a significantly more powerful statement than a "Medium" preference and dramatically more powerful than a "Weak" preference; this position is reflected by the two nonlinear scales. The green curve illustrates a ratio scale with a factor of 2.0, and the blue curve, illustrates a ratio scale with a factor of 3.0. Note on the green curve, a "Strong" (4.0) is valued twice as much as a "Medium" (2.0), and four times as much as a "Weak" (1.0). Correspondingly, observe on the blue curve a "Strong" (9.0) is valued three times as much as a "Medium" (3.0) and nine times as much as a "Weak" (1.0). Choosing a strong blue preference would need to be done deliberately as it carries so much more mathematical weight, which is exactly the point – to emphasize the significance of a 'Strong' preference.

The weakest statement of preference shown in this graph is either of the red scales, while the strongest statement of preference is the blue scale. The specific, numerical scale associated with a preference scale is the choice and privilege of the study designer; however, more important are the arguments (definitions) used to defend the associations' validity. The 0-1-3-9 numerical scale is the foundational association with the qualitative QFD preference scale terms "None", "Weak", "Medium", "Strong" respectively.

Based on this background, the two research questions emerged:

RQ1 – Does the QFD preference scale of 0-1-3-9 need to be as large as a factor of 3.0 in order to reliably distinguish choice?

RQ2 – How small of a difference can human preference reliably perceive?

2 Literature Review

The literature review started by attempting to determine the basis for the TQM QFD 0-1-3-9 weighting scale. When no basis was located, the review moved to the broader subject of human preference where the concept of "Just Noticeable Difference" (JND) and Weber's Law was illuminated (Britt & Nelson, 1976).

2.1 Weber's Law

Definition of Weber's Law and the concept of JND are offered early in the Britt & Nelson (1976) article and translated into a simple proportion: k (constant ratio) = JND/initial stimulus. For example if the weight of an article must be increased 33% before the difference is detectable, then the initial stimulus = weight, and the JND = 1.33*weight – weight. The k-value can be expressed as in equation (1):

$$k = (1.33*\text{weight} - \text{weight})/\text{weight} = .33/1 = .33 \quad (1)$$

The article then demonstrates how the k-value ratio is tested against a preset percentage (typically close to or just above 50%) for consumer preference. The simplicity of the article's presented approach is valuable and can be extended to the previously presented linear (0-1-2-3) and nonlinear preference scales (0-1-2-4, 0-1-3-9) where the corresponding k-values would = 1/.5, 1, and 2 respectively. These k values are much larger than the 0.2 and 0.25 presented in the Britt & Nelson (1976) article.

Weber's (and his mentor Fechner's) mathematical model, originally published by Weber in 1860, deals with the perceived changes in the strength of a stimulus to any of the five senses (Britt & Nelson, 1976) as shown in equations (2) and (3).

$$k = (\text{perceived change } [p\Delta]) = (\text{change in stimulus } [\Delta S]) / (\text{initial stimulus } [S]) \quad (2)$$

$$k = p\Delta = \Delta S/S = JND/S \quad (3)$$

Weber's Law was mathematically derived by Fechner (1966) to account for variances in the sensory acuity of different people. Britt (1975) found JND to be useful, accurate and, computable for all human senses. Weber found a JND was always proportional to the increase in weight; i.e., if the JND in weight from 100 lbs is 5 lbs heavier or lighter, then the JND will increase or decrease with weight at a proportion of 5%. There are alternative views, Diamond (1990) and Krueger (1989) both favor a power law (logarithmic relationship) as another option to qualify preference.

Nearly all the research predicated on the quantitative application of Weber's Law and the JND concept was written over 40 years ago. The more recent studies have still been built on this initial philosophy, but the concept has been tailored to many specific fields. For example, many of the more recent studies deal with how consumers perceived changes in price, color, font, weight, size, auditory stimuli, or other various aspects of specific, marketing strategies (Britt, 1975; Labrecque, 2013). In an unusual application, Crawford (2012) applied Weber's JND to the use of haptic (force) feedback in robotics and the resulting speed and performance of grasping objects. Weber's Law remains as a simple, easy-to-apply view on the relationship between changing stimuli, a JND, and, ultimately, choice.

Weber's Law is frequently quoted with a reliability factor in consumer marketing of 50% (Britt & Nelson, 1976) which seems inordinately low for consistent, dependable results in applications beyond marketing. Reliability of 50% would be unacceptable in aerospace engineering/program development applications where a 5% error (95% reliability) is typically considered the bare minimum of acceptance.

2.2 Quality Function Deployment (QFD) scales

Multiple references concerning the history of QFD and using/optimizing QFD were easily located, starting with the seminal U.S. QFD article by Hauser & Clausing (1988); however, no source was located that indicated any specific QFD scale basis or the rationale associated with the specific selection of QFD's nonlinear 0-1-3-9 scale. An informative article tracing the history of QFD from 1978 to the present (Akao & Mazur, 2003) did not illuminate the scale basis. Burke, et al. (2002); Delano, et.al. (2000); Fiorenzo & Alessandro, (1999); Franceschini & Rossetto (1995); and Franceschini & Rupil (1999) each produced excellent advice for employing and optimizing QFD, but none revealed nor discussed the basis for the preference scale. Thus, the driving motivation for this research was to attempt to understand or, if possible, derive the basis for QFD's 0-1-3-9 preference scale.

While QFD is a powerful and easy-to-use decision analysis tool with diverse applications across an equally diverse industry space, it is not a panacea. Correspondingly, the assertions in the basic principle of Weber's Law and a JND suggest one scale for measuring all perceived differences may not be appropriate; rather, the scale of a JND may change depending on the commodity being measured, the individual being tested, and/or the circumstances of the test.

2.3 General human preference

A commonality shared by many of the reviewed marketing articles was a consumer's JND is tied to the effect a company wants to achieve; therefore, companies have invested to understand qualitatively how different stimuli (color, price, size, utility, precision of fit, online product reviews, etc.) might influence a customer's preference (Chiranjeev & Rajneesh, 2011; Decker & Trusov, 2010; Kumar & Nobel, 2008; Labrecque, 2013; Tantillo, et. al, 1995).

Multiple articles were more mathematically/quantitatively focused such as Britt & Nelson (1976) who dealt with how the equation for a JND might be specifically applied to a marketing preference choice between two similar products. Kwong & Wong (2006) and Smeets & Brenner (2016) dealt with how ratio data affected changes which might be perceived in physical object size. Chiou (2008) studied how deleting common features from two products might affect a consumer's preference for those products. Muunich (2006) focused on proving a theorem that could describe how a consumer chooses a reference point for their personal preferences. Augustin (2009) proposed a reformulating of three, different, stimulus measurements to account for the introduction of empirical data. The tendency of inconsistencies in decision makers' preference was cataloged and measured by Sironen, et.al, (2014). Barzilai (2005) waxed on ratios of preference scales, showing them to be undefined which invalidated another popular decision analysis tool, the Analytic Hierarchy Process (AHP).

2.4 Balancing the reliability of choice with the preference

Takeaways from the literature which are independent of experimental method and were heeded in this research included:

- Being able to test for results with a range of reliability spanning from less than 50% to greater than 95%. Results which are accurate only 50% of the time have less value compared with results that are accurate 95%, 99%, or 99.99% of the time.
- Commonly accepted minimal sample size for social experiments should exceed at least $n = 100$, with no specific upper limit.
- To help ensure the absence of bias, no incentives should be provided for a preference experiment.

2.5 Hypothesis

H_a1 – For a JND to be reliable > 95% of the time, a k = 2 (or a 200% increase from one stimuli level to the next) will be necessary (this k-value correlates directly with the QFD 'factor-of-3' increases in its preference scale: 0-1-3-9).

H_a2 – The larger a given pair-wise comparison is in magnitude (weight or size), the more reliably human preference for the larger (either heavier or larger in size) commodity can be measured.

3 Methodology

3.1 Philosophy

Two primary data collection methods were designed to contrast each other: an in-person, convenience sample taken on the Oklahoma State University (OSU) campus with physical weights and an Amazon M-Turk online survey of random respondents based on pair-wise, visual-size differences of circles. Both of these approaches were designed to elicit a preference response choice when presented with different stimuli. A third collection method via an online survey was also employed with OSU students in the College of Education, Health & Aviation's online research solicitation system – called SONA (htp*s://www.sona-systems.com/).*

The employed survey tools, number of respondents, and number of collected pair-wise comparisons from the respondents were tallied as follows:

- *OSU in-person survey on campus* (*n* = 127 respondents [not paid], 633 comparisons)
- *Amazon's M-Turk* (*n* = 524 respondents [who were paid $0.35 each], 12,048 comparisons)
- *OSU SONA* (n = 210 respondents [not paid], 4,841 comparisons*)*

- Totals: *N* = 861: with 17,522 comparisons

Important note: The overall totals cannot meaningfully be combined as the data streams are fundamentally different – each data stream will be considered individually and compared against the other two streams but not combined.

In addition to the weight data collected, age demographics were also elicited from each respondent according to the following scale shown in Figure 2:

Age:	<22	22-34	35-52	53-71	>71

Fig 2. Age Demographic Scale. The age demographic scale used in this research correlates to generally accepted U.S.-based generational labels respectively (2017 calendar year basis), < 22 = Millennials, 22-34 = Gen Y, 35-52 = Gen X, 53-71 = Baby Boomers, and > 71 = Traditionalists (Center for Generational Kinetics, 2016).

During the in-person survey, age data was collected with a non-attributed slip of paper. In the online version, this request was presented immediately after the consent statement but before the start of the survey.

A collection of desired, pair-wise comparisons was assembled and applied identically in the weight-based or visually-based approach, meaning the same pair-wise comparisons were elicited for presentation in either approach; 23, pair-wise comparisons were available. Three of the 23 pairs were identical and included randomly to complement the 20 dissimilar pairs. A structured approach to the distribution of the weights and shape pairings was used to ensure an equal as possible sampling was gathered among the desired, pair-wise comparisons. In-person respondents were presented with five pairs of weights each, whereas online respondents were presented with 23 pairs of circles of differing areas.

Subjects did not know what weight or shape pairs they were being asked to sample in advance and in both approaches the presentation of the pairs was randomized to prevent the detection of any pattern. A collection of hollow containers filled with varying amounts of ballast was envisioned for presentation to respondents in the in-person version of the survey. Ballast possibilities originally included the consideration of commercially identical shapes, such as marbles or batteries, and measured commodities such as clay, sand, or stones. Due primarily to cost considerations, ¼-20 plated nuts were chosen (as over 650 were needed). A ¼-20 nut is hexagonal, ¼ inch in internal diameter, and threaded at 20 threads per inch. Figure 3 shows a small collection of ¼-20 plated nuts cast against a 6-inch ruler.

Fig 3. Ballast Choice. ¼-20 plated nuts were chosen as the ballast for the research because

they are small, uniform, and inexpensive compared with considered alternatives.

The following paragraphs offer more detail about the research script, sample size, and the contrasting data collection approaches of in-person versus online.

3.2 Research Script

Table 1 shows on the left side two perspectives for needed-weight pairs. The top row is one perspective, while the bottom three rows show the second perspective. The far right three columns show the cumulative number of ballast articles (the ¼-20 plated nuts) needed – which totaled 605.

Table 1. Experimentation Pair-Wise Comparison Script

10% Increase Basis			33% Increase Basis			50% Increase Basis			100% Increase Basis		
smaller	med	larger	smaller	med	larger	smaller	med	larger	smaller	med	larger
10-11			3-4			2-3			1-2		
	20-22			15-20			10-15			10-20	
		40-44			30-40			20-30			20-40

	Smaller			Medium			Larger		
Linear Increase (x1) Basis (100 % / 50%)	1	2	3	3	6	9	9	18	27
	1-2			3-6			9-18 (10-20)		
		2-3			6-9			18-27 (20-30)	
Non-Linear Increase (x2) Basis (100%)	1	2	4	3	6	12	9	18	36
	1-2			3-6			9-18 (10-20)		
		2-4			6-12			18-36 (20-40)	
Non-Linear Increase (x3) Basis (300%)	1	3	9	3	9	27	9	27	81
	1-3			3-9			9-27		
		3-9			9-27			27-81	

Quantity	#	Totals
1	2	2
2	2	4
3	4	12
4	2	8
6	3	18
9	3	27
10	3	30
11	1	11
12	1	12
15	2	30
20	5	100
22	1	22
27	2	54
30	1	30
40	3	120
44	1	44
81	1	81
		605

Note: This table shows, in the yellow boxes, the 20 numerical pairs associated with the experimental, pair-wise comparisons used in the research. The top row of the table shows one approach (10/33/50/100% increase), while the three lower rows show a multiplicative factor of x1, x2 or x3 weight increase approach. The vertical tally on the right side shows the total number of ballast articles needed to complete the experiments at 605, which are distributed among the yellow-box pairings shown to the left. For example the top, left yellow box shows a 10-11 pair-wise comparison, meaning one sample presented to a respondent contained 10 ballast weights while the corresponding sample contained 11 ballast weights. The total of the needed ballast in the 20 yellow boxes = 605.

The first perspective on structuring pair-wise weights is represented in the top row of the table by the 10%, 33%, 50%, and 100% increase basis boxes, each of which contains a "Smaller", "Medium", and "Larger" quantity. The starting point for each basis was the smallest number of ¼-

20 plated nuts which would be needed to build that basis in whole units. Partial ¼-20 plated nuts were not allowed. A minimum of a doubling was desired moving from the "Smaller" to "Medium", and again from "Medium" to "Larger".

The second perspective stems from the QFD scale basis, presented in the beginning of the paper in Figure 1, and is shown in the bottom three rows of the table. The first row of this section started with a basic, linear (x1 unit) increase, proceeded to a nonlinear (x2 unit) increase basis and concluded with a nonlinear (x3 unit) increase identical to the QFD 0-1-3-9 preference scale. Each basis (x1, x2, and x3) was scheduled to be incremented in a similar "Small", "Medium" and "Larger" quantity fashion.

In the first row of this lower section, a 1 is the smallest available starting point, thus the "Smaller" scale proceeds by 1 unit (x1) increments to 2 and 3. The "Medium" scale starts with a 3 and adds increments of 3 (1 unit) to conclude with a 9. Similarly, the "Larger" scale starts with a 9 and adds increments of 9 (1 unit) to conclude with a 27. This was done again in a similar format for both nonlinear scales on the 2-unit (x2) and 3-unit (x3) basis shown in the bottom pair of rows.

The 20, unique, pair-wise comparisons are highlighted in yellow. The 9-18, 18-27, and 18-36 pairs were adjudged to be adequately represented and indistinguishable from the 10-20, 20-30, and 20-40 pairs respectively in experimentation, so neither the 9-18, 18-27, nor 18-36 were explicitly included in the research script; however, 10-20, 20-30, and 20-40 were included in the research script.

Finally, in addition to these 20 pair-wise comparisons, three additional pairs were added (1-1, 2-2, and 20-20) to explore whether identical pairs of differing weights could accurately be distinguished from similarly weighted, non-equal pairs. This gave a research script of 23 different, pair-wise comparisons for exploration and required 46 additional ballast articles for a grand total of 651 ¼-20 nuts.

3.3 Sample Size

The minimum, desired, pair-wise sample size was determined as follows: $N = (Z^2pq)/(e^2)$. This approach was presented in Barlett, Kotrlik, & Higgins (2001) for categorical data with a Z-estimation at 95%/α = 0.05, Z = 1.96; p = proportional variable of interest = q remaining proportional variable = 0.5 (because N is maximized with both p and q = 0.5), and e = acceptable error = 0.05, then N = 384.

Each of the three survey methods exceeded N=384. For the in-person survey, in which each respondent was presented with five pairs of boxes, this required a minimum of n = 384/5 = 77 (rounded up) respondents to obtain the required 384 pair-wise comparisons. From 127 respondents, 633 pair-wise comparisons were obtained, not the expected 127 * 5 = 635 because one respondent did not complete the requested five-pair sampling, rather they completed three.

The online M-Turk survey, which contained 23 pair-wise comparisons, required n = 384/23 = 17 respondents; however, given the expected ease of collecting this data and the minimal time necessary to assess the comparisons, over 500 respondents were obtained in a less than two-day window, returning 12,048 pair-wise comparisons. The OSU SONA collection enjoyed more than n = 200 respondents, offering a further 4,841 pair-wise comparisons.

3.4 In-Person Survey (OSU campus)

The 46 individual collections of ¼-20 plated nuts comprising the 23 pairs were packaged in commercially obtained, identical-sized, white, sealed, cardboard boxes measuring approximately 2.0 x 2.0 x 4.0 inches in exterior dimensions. Each box was alphabetically coded in the same spot so the researches knew the contents. The nuts were secured in the boxes with packing tape so movement of the box would not reveal its contents. Figure 4 shows a folded box (labeled "A") which contained 10 ¼-20 plated nuts on top of a new, unfolded box.

Fig 4. Respondent Sampling Boxes. This is one of the 23 boxes prepared for the research on top of a stock, new, unfolded box (note the small "A" inscription on the top edge). This alphabetical code, which only faced the researchers, indicated the box contained 10 ¼-20 nuts.

A methodological flaw the researchers noticed was the ¼-20 nuts were not always secured to the same surface within the box; this potentially introduced experimental bias as some boxes had ¼-20 nuts secured to the rear wall of the box instead of the bottom. These afflicted boxes had a "natural" center-of-gravity rotation moment when picked up, compared with the boxes where the ¼-20 nuts were affixed on the bottom had no rotation moment.

A table with an official OSU Aviation Program tablecloth was set up (with appropriate university permission) adjacent to the student union on four, different days in April 2017. To lend additional authenticity to the in-person research collection, each of the researchers (two students and the principal investigator) were "uniformed" with the same OSU Aviation Program shirts. All 46 preassembled boxes were arrayed in a single line from left to right in an order not discernable to the respondents as the coded annotations were facing the researchers. As passersby's approached, they were verbally encouraged to stop for a few moments and participate in the research project. Once they had agreed, the following statement was read to each respondent:

We are conducting preference research with the goal of understanding how humans differentiate between choices. We are using small weights to explore choice. You are welcome to review our recruitment statement as it entails your rights in this experiment and the procedure we'll be following. Would you like to read it (offered, if requested)? Before we begin, this is by no means required but if you are comfortable, would you mind indicating your age group on this slip? We are going to place five

pairs of boxes in front of you. Then we will ask you to pick them up and answer the principal question, "Can you detect a difference in weight? If so, please describe the difference?" You can respond any way you wish such as "Heavier", "Lighter", or "Equal". After you complete the comparisons, please feel free to take a few lollipops as our thank you for your time."

After agreeing to participate, the respondent was offered an age-demographic survey to confidentially circle their appropriate age bracket. This survey was then placed in a sealed envelope.

The 23 box pairs were pre-subdivided into groups of five to ensure each pair was sampled as near equally as possible. One pair at a time, each of the five pairs of boxes was then randomly selected by a researcher and placed on the table in front of the respondent with the coded annotation facing the researcher. The respondent was allowed as much time as they desired to make their comparison and determine a weight difference. The required oral responses were on the order of: "Left is heavier/lighter", "Right is heavier/lighter", or "Both are the same". Rarely did the respondent assessment take more than 3-to-5 seconds per pair-wise comparison. Once completed, the sampled boxes were collected by the researcher and returned to the array of 46 boxes on the table and the process repeated with a new pair of boxes. At any time, only one pair of boxes was available for the respondent to handle. The entire visit with each respondent lasted about 3-4 minutes.

3.5 Online Survey (Amazon M-Turk / OSU SONA)

This form of the survey compared the size of two, side-by-side circles whose areas had been calculated to respectively match each one of the 23 pairs of ¼-20 nut quantities (weights) prescribed in the experimental script. Identical to the in-person survey, the respondent was offered the choices of "Left", "Equal", or "Right" to indicate which circle they perceived to be larger.

Consistent with Krider, Raghubir, & Krishna (2001), area was selected as the visual metric to be scaled as it proportionally represented total weight more accurately than either a proportional increase in diameter or circumference. For example, if the script called for a 10-20 pairing of ¼-20 nuts, which equated to a k factor of 1.0, or a 100% increase from 10-to-20, this was easily accomplished with the boxes and ¼-20 nuts but visually, three choices were possible: a) increase the diameter from 10-to-20, b) increase the circumference from 10-to-20, or c) increase the area from 10-to-20. Only choice c) preserves a k = 1.0 perspective. Figure 5 shows these relationships drawn to scale. Note both a diameter and circumference percent increase represent an area increase of 300% or a k = 3.0, not the desired k = 1.0. The size of the circles in the bottom row of this figure is exactly how an online respondent would have seen the 10-to-20 relationship presented to them electronically. None of the verbiage shown below in Figure 3 was visible when circles were being presented to the respondents.

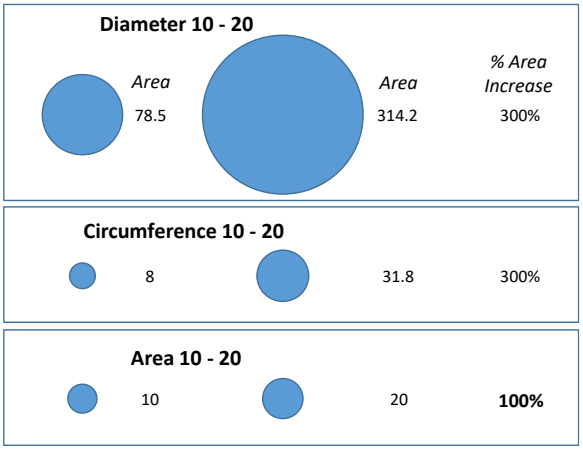

Fig 5. Selecting Circle Area. These three, scaled plots show the relative size differences in area diameter, circumference, and area when a 10-to-20 ratio respectively is used in a percent increase computation. Each row starts on the left with a constant basis of 10; e.g., in the diameter row, the left circle has a diameter of 10 (units) compared to the right circle with a diameter of 20 (units). The corresponding areas associated with those dimensions are shown and even though the diameter doubles, the area quadruples, thus a 300% increase. The same relationship is evident in the middle row where the left circle has a circumference of 10 (units) and grows to a circumference of 20 (units) in the right circle; which is a doubling of circumference, but the area again quadruples. The last row shows a purposefully scaled circle with an area of 10 (units) growing to a circle with an area of 20 (units), a doubling of desired area, thus a 100% increase.

Qualtrics® was used to construct and deliver both online surveys. The 23 circle pairs were proportioned identically (normalized) in area from the ¼-20 nut pair weights. The 23 pairs were randomized in their order of presentation, and there were two versions of each circle pair, one with the larger circle on the right and the other with the larger on the left. In addition to the random order, the left or right orientation of the larger circle was also randomized equally in the survey presentation. A unique URL was added to distinguish the M-Turk data feeds from the OSU SONA data feeds. Figure 6 shows the online "sample" screen (with a purposefully exaggerated k > 3.0) and the immediately following selection screen.

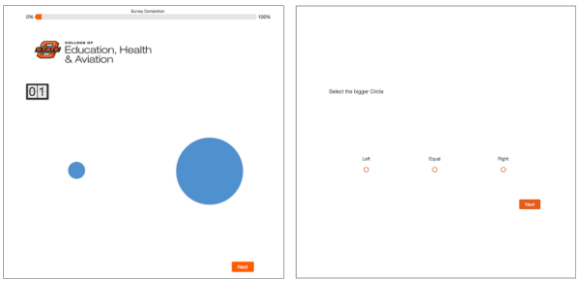

Fig 6. Respondent Circle Sampling. The left panel of this figure shows the "sample" circle pairing shown to respondents in the online survey prior to data collection. The distance between the centers of the circles was constant, only the sizes of the circles changed from one elicitation to the next. The "01" 'flip box' in the upper left was a counter (which started at "03" and counted backwards to "00") to provide 3 seconds of viewing time. Qualtrics® then automatically sequenced the presentation to the right panel to collect the respondents selection of which circle was larger. The positioning of the clickable selection buttons in the right panel was purposefully placed to match the centers of the preceding circles.

Figure 7 Shows the online-screen presentation of a k =.1 iteration sized to replicate the 40-to-44 ¼-20 nut pair. This k = .1 size increase in area from left to right is subtle and was frequently scored incorrectly as "Equal".

Fig 7. Online Respondent Presentation. This screen shot shows the online presentation of the 40-to-44 ¼-20 nut pair translated (normalized) to area. Note the counter has incremented to 2 seconds and the orange slider bar at the top of the screen is showing ~ 25% complete with the survey, so this would likely have been the sixth circle pair shown to the respondent.

The circle pairs were shown to the respondent for a maximum of three seconds, then the Qualtrics® survey tool advanced automatically to the next screen on which the circles were removed and three clickable choices were presented in the same spot equidistantly on each respective solicitation screen from left-to-right as follows: "Left", "Equal", "Right".

The respondents could take (theoretically) as much time as desired to make their choice before advancing to the next pair-wise comparison, but once advanced, the three-second counter started again. Beta tests of the online version revealed the 23 pairings could easily and accurately be sampled in a 3-to-5 minute window. Except for the unique URL associated with each online Qualtrics® survey, the online survey tool was identical in both the M-Turk and OSU SONA applications.

4 Results

Analysis and interpretation of the results sought an answer to both the presented research questions and hypotheses. Pursuant to the methodological statement previously made, the three, data streams (1) OSU in-person (heretofore referred to as the collection mode "¼-20 Nuts"), and the two online surveys, (2) M-Turk, and (3) OSU SONA are shown separately and compared but never combined.

The data streams differed fundamentally in the following ways and therefore were not combined:

- Methodological basis: weight v. area and the associated sense, touch v. sight.
- The OSU in-person solicitation was a convenience survey, whereas both the M-Turk and OSU SONA data were randomly sampled.
- The M-Turk data showed significantly more age diversity than either of the OSU surveys, but all of the surveys were biased toward young adults when compared with the U.S. population.
- The OSU SONA data also represented a nearly homogenous sample from, very likely, the same population as the OSU in-person survey.
- While it is possible, and arguably reasonable, to combine the two OSU data streams, the researchers chose not to because they were unsure in the online version whether the selection of the "Right" clickable button in the selection screen was unduly influenced by its proximity to the "Next" button, making it too convenient to accelerate the survey by oscillating between these two buttons. This same concern was also applicable to the M-Turk survey and while neither data stream results confirm this bias, the likelihood was adjudged more probable in the collegiate student population.

All data collected via the three data streams and the conversions used between the weights (¼-20 nuts)/areas of the respective circles is contained and displayed in the Appendix A tables.

Figure 8 presents a macro view of the research results, cataloged by data stream. Each dot is a visual representation of one of the 23 rows of data contained in each of Appendix Tables A1 (in-person survey [blue]), A3 (M-Turk survey [orange]), and A4 (SONA survey [grey]). There are thus 69 dots in the figure, 23 of each respective color; however, due to the order in which data is selected for plotting, MSExcel will plot on top of previous points if subsequent points are the same value. All 23 SONA data points (grey) are visible because they were selected for plotting last. Only 17 of the in-person/¼-20 Nuts data points (blue) are visible because they were selected for plotting first.

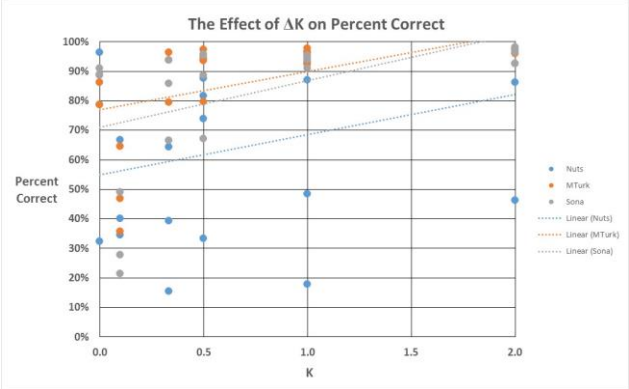

Fig 8. Summary of Collected Data by k-factor. All collected survey data is shown in this figure, with the three collection streams color-coded. The k-value is incremented according to the survey script from k = 0, .1, .33, .5, 1.0, to 2.0 along the horizontal axis. The vertical axis shows the percent of correct responses, thus the reliability of the responses. The online data streams (M-Turk [orange], and SONA [grey]) show stronger and similar convergence compared with the in-person (OSU survey [blue]). Trend lines for all surveys were positive and of similar slope.

The trend for all data was as the k-value increases, the reliability also increased. There was significant reliability diversity within the in-person data stream and for all data streams below k = 1.0, especially below k =.5.

Figure 9 breaks Figure 8 into its three respective, data stream components and shows each uniquely. This view highlights both the scatter of the OSU in-person survey data (¼-20 Nuts) and the rapid convergence of the online data streams (M-Turk and SONA).

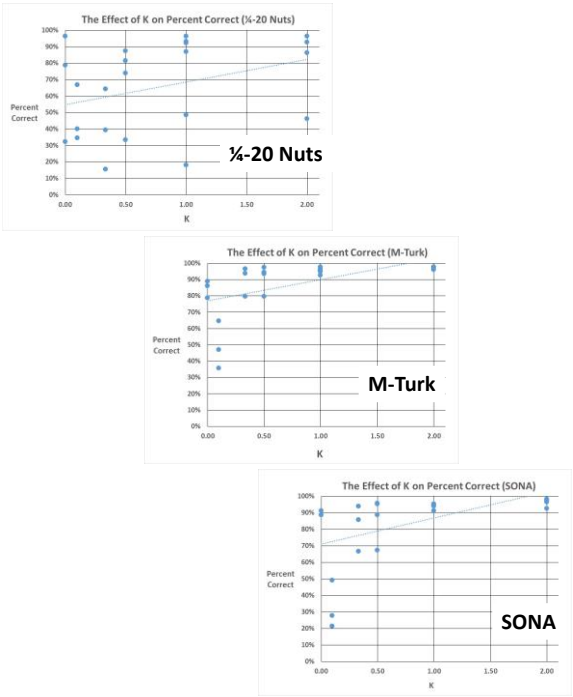

Fig 9. Separate Data Streams. Each of the three data streams is shown individually: the top graph illustrating the OSU in-person data; the middle graph, the M-Turk online data; the bottom graph, the OSU SONA online data. While the slope of the trend lines are nearly identical, their placement is not. The OSU in-person survey data shows more variability and less overall reliability than either of the online collection streams. All three data streams show high and low reliability at small k-values, especially below k = .5. Universally, the data points in the lower left of the graphs reflect the smallest weight differences.

Observations from Figure 9 include:

- In-person, weight-based survey produced inferior reliability results to either online, area-based elicitation.
- The online results were remarkably consistent, which was potentially significant considering they were drawn from different populations.
- k = .1 was confounding to respondents no matter whether weight-based or area-based. Note in the online surveys, the k = .1 reliability was inferior to all other k-values, even when k = 0.
- When k = 0 (equal weights or equal areas), increasing the weight/area difference drove reliability into retrograde, contrary to what was observed in every other weight/area pair case. This was most aggressively noticeable in the OSU in-person, weight-based survey; reliability for the 1-1 pair was 96%, the 2-2 pair fell to 79%, and the 20-20 pair plummeted to 32%. The trend was similar but less dramatic for the M-Turk sampling spanning from 89% to 79% reliability and confounded in the OSU SONA online survey, vacillating around +/- 1% around 90%.
- Unexpectedly, when the magnitudes of the weights (or areas) were very small, it was easier for respondents to accurately sense both weights (and areas) were equal. As the magnitude of the weights (or areas) became larger, respondent's ability to accurately determine equality degraded and especially with the ¼-20 nut weights (as shown in the Appendix A data tables). Correspondingly it was consistent no matter the data source, approximately 1-out-of-100 respondents chose from the 9-27 or 27-81 pairs either the weight or area was three times heavier or larger and stated it was the "Lighter" or "Smaller", respectively.
- The OSU in-person survey data was more scattered, especially at lower weight/area pairs and at all k-values.
- The rate-of-reliability increase was more significant with the OSU in-person survey data across all k-values > 0; whereas, the rate of reliability increase was shallower with either online survey method for k-values of .33 and .5 and was consistently shallow (flat) with online k-values of 1.0 and 2.0.

A Chi-square test for independence was conducted between the comparable data streams. To conduct this test, the focus was the number of pair-wise comparisons that achieved a 95% or greater reliability. The first test compared the OSU survey streams (in-person and SONA), and the second test compared the two, online survey sources (M-Turk and OSU SONA). Because they came from the same population, it would have been logical to suspect the OSU survey data should be related. Table 2 shows the number of pair-wise comparisons which met the 95% reliability threshold (reference Appendix tables A1 and A4 for the actual data).

Table 2. Chi-square data comparing OSU survey sources

	<95%	>=95%	Totals
OSU in-person	20	3	23
OSU SONA	14	9	23
	34	12	**46**

The statistically significant result was the two sets of scores were dependent, which was expected because $\chi^2obt = \sum(f_o - f_e)^2/f_e = 4.059$, which exceeds χ^2crit at 1 degree of freedom ($\alpha = 0.05$) = 3.841.

A Chi-square test for independence was also conducted with the same 95% reliability threshold between the two, online survey streams as shown in Table 3 (reference Appendix tables A3 and A4 for the actual data).

Table 3. Chi-square data comparing online survey sources: M-Turk and OSU SONA

	<95%	>=95%	Totals
M-Turk	13	10	23
OSU SONA	14	9	23
	27	19	**46**

The statistically significant result was the two sets of scores were independent because $\chi^2obt = 0.090$, was less than χ^2crit at 1 degree of freedom ($\alpha = 0.05$) = 3.841. This result helped establish the case that the OSU SONA data was directly comparable with the M-Turk data, in spite of the fact the two surveys drew from very different populations.

Figure 10 changes the x-axis presentation perspective shown in Figures 8 and 9 from k to the mean difference in either the weight or area pair for a specific k. Recall the circle areas were normalized to follow the exact mathematical relationships scheduled for the ¼-20 nuts. For example, to show the results by weight (or area) for the 27-81 pair (k = 2.0), the mean of this pair is computed as (27 + 81)/2 = 108/2 = 54. Data for the 27-81 pair was plotted on the horizontal axis at 54, not 27, or 81. Figure 10 organizes the mean weight/area pairs by constant factor (a multiplier) of 1x, 2x, or 3x. These graphs were assembled to directly contrast linear weight increases (factor of 1x increase) with the nonlinear increases (factors of either 2x or 3x increase).

The three, equal, pair-wise comparisons (1-1, 2-2, 20-20) were not included in three Figure 10 graphs; thus, there were a total of only 20 data points for each survey stream shown, not 23 data points per survey stream as previously shown in Figures 8 and 9.

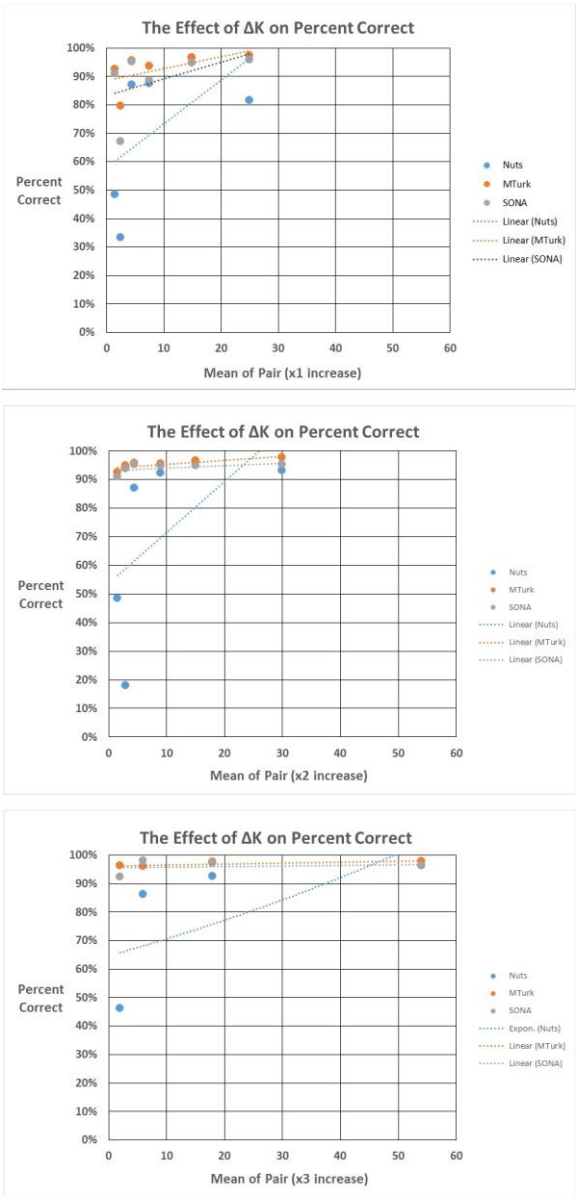

Fig 10. Summary of Collected Data by Weight. These three graphs show on the x-axis by mean weight/area, the increase in pairs selected at a 1x, 2x, and 3x factor and the resulting effect on reliability on the y-axis. All three data streams are shown in respective color codes. Each factor level was composed of the "Smaller", "Medium", and "Larger" set of pairs as shown in the lower three rows of Table 1.

Predictably, when the weight/area is increased at a larger rate (3x v. 2x, or 1x), the reliability increases quickly to over 90% and, in many cases for the online survey data, over 95% even at low weight differences. These graphs again show the area-based results produced superior reliability than the weight-based results. All of the OSU in-person survey results [blue] scoring less than 50% reliability are for small weight differences (1 or 2 ¼-20 Nuts).

4.1 Age Data

The original intent in collecting an age demographic was to associate age with reliability. This required every respondent's pair-wise comparison selections be associated with their age. During the initial, in-person elicitation this quickly became an untenable, data-collection, bookkeeping challenge and was abandoned in favor of simply asking the respondent to offer their age bracket. Age data was successfully collected for most, but not all, respondents in the OSU surveys and was collected from all respondents in the M-Turk online survey. Table 4 shows the age demographic data.

Table 4. Surveyed Age Demographic Data

Age:	<22	22-34	35-52	53-71	>71	Totals
OSU in-person	78 - 64%	35 - 29%	2 - 16%	7 - 6%	0 - 0%	122 - 100%
M-Turk	20 - 3%	285 - 50%	190 - 33%	78 - 14%	0 - 0%	573 - 100%
OSU SONA	136 - 66%	63 -31%	7 -3%	0 - 0%	0 - 0%	206 - 100%

Note: For each survey source, the raw number of respondents in the respective age bracket is followed by its respective percentage of the total.

None of the survey sources mirror general, U.S., 2010 census (2011) data nor is it necessarily reasonable to expect they should. Predictably, the OSU sources were very similar, especially in the lower two age brackets. The M-Turk data is significantly more dispersed than either the OSU in-person survey or the OSU SONA survey, but it too is still disproportionately youthful compared to the U.S. population. College campuses and those willing to complete online surveys for very small sums of compensation appear to appeal to a more youthful demographic. 2010 U.S. census data shows the 18-24 age group = 9.6% of the population, 25-44 = 30.2%, 45-64 = 22%, and 65 and older = 12.4%. Both sampled populations in this research are distinctly younger.

5 Summary

5.1 Observation

- This research compared the OSU population demographic with a similar, youthful, U.S. population demographic. While the overall survey results cannot statistically be claimed to represent the U.S. general population, the strong similarity in the M-Turk demographics with the U.S. census demographics suggests the observed preference trends in the online surveys could be expected in the younger, U.S. general population age demographics.
- There were minor potential biases unintentionally induced in the data collection; if performed again, they would be easy to correct (standardizing the position of the ballast in the sampling boxes and placing the online "Next" sequencing button in a neutral position equidistant from any offered choice button).
- Visual-based, preference data appears significantly more reliable than weight-based, preference data.
- To achieve high reliability (defined as > 95%) in preference, the JND needs to be at least a factor of 3 (k = 2.0) between the choices – in either weight or viewable area.
- Online data collection in this research was significantly faster, took less time to administer, and produced larger data streams.
- Assessing weight with tactile, human touch would be impossible to do online and, if important, must be gathered in person. Assessing weight could be done remotely with robotic force-feedback controls but would still require a longer interface time for each assessment compared with an online, visual-based assessment.
- Future research possibilities include exploring whether the same visual (area-based) results would be replicated in-person using paper diagrams and expanding the data collection to other human senses such as hearing, smell, taste, or other forms of touch (e.g. temperature)

5.2 Propositions to Research Questions/Hypotheses:

RQ1 – Does the QFD preference scale of 0-1-3-9 need to be as large as a factor of 3.0 in order to reliably distinguish choice? – Yes, this research confirmed if 95% reliability was desired, then the perceived difference needed to be a factor of 3.0.

RQ2 – How small of a difference can human preference reliably distinguish? – This research found very small differences were reliably detected across both weight/area elicitations if a 50% accuracy level was acceptable. Weight/area differences as small as 10% (k = .1) were reliably detected with 50% accuracy with larger magnitude weight pairs, or scaled areas of the same ratio.

If the accuracy required was 95% then only two weight difference pairs were able to achieve this level, independent of the sampling method (10-20 [k = 1.0] and 27-81 [k = 2.0]). Multiple instances of 95% accuracy were achieved with the area-based approach, starting at much lower k-values of k = .33.

H_a1 – For a JND to be reliable > 95% of the time, a k = 2.0 (or a 200% increase from one stimuli level to the next) would be necessary (this k-value correlates directly with the QFD "factor-of-3" increases in its preference scale: 0-1-3-9). – Yes, this was the minimum level of k-value tested which produced this level of 95% accuracy. – Reject the Null, accept the hypothesis.

H_a2 – The larger a given pair-wise comparison is in magnitude (weight or size), the more reliably human preference for the larger (either heavier or larger in size) commodity can be measured. – Yes all three streams of survey data exhibited this behavior. – Reject the Null, accept the hypothesis.

Appendix A

Table A1. OSU In-Person Survey Data

OSU Convenience Preference Survey Responses (¼-20 nuts):													
Pair:	**Left:**	**Label**	**Right:**	**Label**	**Totals**	**Left:**	**Right:**	**Equal:**	**Factor**	**L&R Mean**	**k**	**% Increase**	**% Correct**
1	1	H	1	H	27	0	1	26	1.00	1.00	0.00	0%	96.3%
2	2	G	2	G	28	3	3	22	1.00	2.00	0.00	0%	78.6%
3	20	B	20	B	31	7	14	10	1.00	20.00	0.00	0%	32.3%
4	10	A	11	AA	25	4	10	11	1.10	10.50	0.10	10%	40.0%
5	20	B	22	BB	32	12	11	9	1.10	21.00	0.10	10%	34.4%
6	40	C	44	CC	30	4	20	6	1.10	42.00	0.10	10%	66.7%
7	3	D	4	DD	26	8	4	14	1.33	3.50	0.33	33%	15.4%
8	15	E	20	EE	28	1	11	16	1.33	17.50	0.33	33%	39.3%
9	30	F	40	C	28	2	18	8	1.33	35.00	0.33	33%	64.3%
10	2	G	3	D	30	5	10	15	1.50	2.50	0.50	50%	33.3%
11	6	I	9	J	24	2	21	1	1.50	7.50	0.50	50%	87.5%
12	10	A	15	E	23	3	17	3	1.50	12.50	0.50	50%	73.9%
13	20	B	30	F	27	0	22	5	1.50	25.00	0.50	50%	81.5%
14	1	H	2	G	31	3	15	13	2.00	1.50	1.00	100%	48.4%
15	2	G	4	DD	28	3	5	20	2.00	3.00	1.00	100%	17.9%
16	3	D	6	I	23	1	20	2	2.00	4.50	1.00	100%	87.0%
17	6	I	12	K	26	1	24	1	2.00	9.00	1.00	100%	92.3%
18	10	A	20	B	28	0	27	1	2.00	15.00	1.00	100%	96.4%
19	20	B	40	C	29	1	27	1	2.00	30.00	1.00	100%	93.1%
20	1	H	3	D	26	3	12	11	3.00	2.00	2.00	200%	46.2%
21	3	D	9	J	29	1	25	3	3.00	6.00	2.00	200%	86.2%
22	9	J	27	L	27	1	25	1	3.00	18.00	2.00	200%	92.6%
23	27	L	81	M	27	0	26	1	3.00	54.00	2.00	200%	96.3%
				total:	633								
				avg:	27.5								

Note: This table and the corresponding tables for the M-Turk and SONA data catalog the results according to the script shown in Table 1. Each table is organized by k-value then by weight (area) within the respective k-values. The tables do not reflect the randomized deployment of the survey data in either the order of the pairs nor the randomized left-right orientation of the larger weight/area. The "Left" and "Right" labels refer to the alphabetical codes affixed to the reverse side of the small boxes presented for evaluation to the respondents. These codes allowed the research team to discretely keep track of the box contents without biasing the respondents. The red "% Correct" cells in the far right column are unexpected discontinuities in the data and have no plausible explanation.

Table A2. Conversion from Weight to Area for all 23 Tested Pairs

% Area Increase						
Pair:	**Left Diameter**	**Left Area**	**k**	**% Increase**	**Right Area**	**Right Diameter**
1	**1.0**	0.8	0.00	0%	0.8	**1.0**
2	**2.0**	3.1	0.00	0%	3.1	**2.0**
3	**20.0**	314.2	0.00	0%	314.2	**20.0**
4	**10.0**	78.5	0.10	10%	86.4	**10.5**
5	**20.0**	314.2	0.10	10%	345.6	**21.0**
6	**40.0**	1,256.6	0.10	10%	1,382.3	**42.0**
7	**3.0**	7.1	0.33	33%	9.4	**3.5**
8	**15.0**	176.7	0.33	33%	235.6	**17.3**
9	**30.0**	706.9	0.33	33%	942.5	**34.6**
10	**2.0**	3.1	0.50	50%	4.7	**2.4**
11	**6.0**	28.3	0.50	50%	42.4	**7.3**
12	**10.0**	78.5	0.50	50%	117.8	**12.2**
13	**20.0**	314.2	0.50	50%	471.2	**24.5**
14	**1.0**	0.8	1.00	100%	1.6	**1.4**
15	**2.0**	3.1	1.00	100%	6.3	**2.8**
16	**3.0**	7.1	1.00	100%	14.1	**4.2**
17	**6.0**	28.3	1.00	100%	56.5	**8.5**
18	**10.0**	78.5	1.00	100%	157.1	**14.1**
19	**20.0**	314.2	1.00	100%	628.3	**28.3**
20	**1.0**	0.8	2.00	200%	2.4	**1.7**
21	**3.0**	7.1	2.00	200%	21.2	**5.2**
22	**9.0**	63.6	2.00	200%	190.9	**15.6**
23	**27.0**	572.6	2.00	200%	1,717.7	**46.8**

Note: This table is also organized by k-value. No units were shown to respondents as all circles appeared without scale. All circles were drawn to a constant scale based on diameter.

Table A3. M-Turk Online Survey Data

M-Turk Preference Survey Responses (Online):													
Pair:	**Left:**	**Label**	**Right:**	**Label**	**Totals**	**Left:**	**Right:**	**Equal:**	**Factor**	**L&R Mean**	**k**	**% Increase**	**% Correct**
1	0.0080	H	0.0080	H	525	22	37	466	1.00	1.00	0.00	0%	88.8%
2	0.0310	G	0.0310	G	524	18	55	451	1.00	2.00	0.00	0%	86.1%
3	3.1420	B	3.1420	B	524	59	53	412	1.00	20.00	0.00	0%	78.6%
4	0.7854	A	0.8639	AA	525	16	187	322	1.10	10.50	0.10	10%	35.6%
5	3.1416	B	3.4558	BB	525	9	246	270	1.10	21.00	0.10	10%	46.9%
6	12.5664	C	13.8230	CC	524	13	338	173	1.10	42.00	0.10	10%	64.5%
7	0.0707	D	0.0942	DD	521	6	414	101	1.33	3.50	0.33	33%	79.5%
8	1.7671	E	2.3562	EE	523	4	490	29	1.33	17.50	0.33	33%	93.7%
9	7.0686	F	9.4248	C	521	8	502	11	1.33	35.00	0.33	33%	96.4%
10	0.0314	G	0.0471	D	522	9	415	98	1.50	2.50	0.50	50%	79.5%
11	0.2827	I	0.4241	J	528	7	494	27	1.50	7.50	0.50	50%	93.6%
12	0.7854	A	1.1781	E	524	11	494	19	1.50	12.50	0.50	50%	94.3%
13	3.1416	B	4.7124	F	525	10	511	4	1.50	25.00	0.50	50%	97.3%
14	0.0079	H	0.0157	G	524	8	485	31	2.00	1.50	1.00	100%	92.6%
15	0.0314	G	0.0628	DD	521	6	494	21	2.00	3.00	1.00	100%	94.8%
16	0.0707	D	0.1414	I	525	10	502	13	2.00	4.50	1.00	100%	95.6%
17	0.2827	I	0.5655	K	524	10	500	14	2.00	9.00	1.00	100%	95.4%
18	0.7854	A	1.5708	B	523	10	505	8	2.00	15.00	1.00	100%	96.6%
19	3.1416	B	6.2832	C	524	6	512	6	2.00	30.00	1.00	100%	97.7%
20	0.0079	H	0.0236	D	528	7	508	13	3.00	2.00	2.00	200%	96.2%
21	0.0707	D	0.2121	J	522	5	501	16	3.00	6.00	2.00	200%	96.0%
22	0.6362	J	1.9085	L	525	5	512	8	3.00	18.00	2.00	200%	97.5%
23	5.7256	L	17.1767	M	521	6	509	6	3.00	54.00	2.00	200%	97.7%
				total:	12048								
				avg:	523.8								

Table A4. OSU SONA Online Survey Data

OSU SONA Preference Survey Responses (Online):													
Pair:	**Left:**	**Label**	**Right:**	**Label**	**Totals**	**Left:**	**Right:**	**Equal:**	**Factor**	**L&R Mean**	**k**	**% Increase**	**% Correct**
1	0.0080	H	0.0080	H	212	6	18	188	1.00	1.00	0.00	0%	88.7%
2	0.0310	G	0.0310	G	211	2	17	192	1.00	2.00	0.00	0%	91.0%
3	3.1420	B	3.1420	B	211	12	12	187	1.00	20.00	0.00	0%	88.6%
4	0.7854	A	0.8639	AA	210	6	45	159	1.10	10.50	0.10	10%	21.4%
5	3.1416	B	3.4558	BB	210	4	58	148	1.10	21.00	0.10	10%	27.6%
6	12.5664	C	13.8230	CC	210	3	103	104	1.10	42.00	0.10	10%	49.0%
7	0.0707	D	0.0942	DD	212	4	141	67	1.33	3.50	0.33	33%	66.5%
8	1.7671	E	2.3562	EE	210	2	180	28	1.33	17.50	0.33	33%	85.7%
9	7.0686	F	9.4248	C	209	1	196	12	1.33	35.00	0.33	33%	93.8%
10	0.0314	G	0.0471	D	210	3	141	66	1.50	2.50	0.50	50%	67.1%
11	0.2827	I	0.4241	J	209	6	185	18	1.50	7.50	0.50	50%	88.5%
12	0.7854	A	1.1781	E	210	1	200	9	1.50	12.50	0.50	50%	95.2%
13	3.1416	B	4.7124	F	210	3	201	6	1.50	25.00	0.50	50%	95.7%
14	0.0079	H	0.0157	G	211	5	192	14	2.00	1.50	1.00	100%	91.0%
15	0.0314	G	0.0628	DD	210	3	197	10	2.00	3.00	1.00	100%	93.8%
16	0.0707	D	0.1414	I	209	2	199	8	2.00	4.50	1.00	100%	95.2%
17	0.2827	I	0.5655	K	211	5	200	6	2.00	9.00	1.00	100%	94.8%
18	0.7854	A	1.5708	B	211	3	200	8	2.00	15.00	1.00	100%	94.8%
19	3.1416	B	6.2832	C	211	4	201	6	2.00	30.00	1.00	100%	95.3%
20	0.0079	H	0.0236	D	211	3	195	13	3.00	2.00	2.00	200%	92.4%
21	0.0707	D	0.2121	J	211	2	207	2	3.00	6.00	2.00	200%	98.1%
22	0.6362	J	1.9085	L	211	2	205	4	3.00	18.00	2.00	200%	97.2%
23	5.7256	L	17.1767	M	211	3	203	5	3.00	54.00	2.00	200%	96.2%
				total:	4841								
				avg:	210.5								

References

Akao, Y., Mazur, G.H. (2003). The leading edge in QFD past, present and future, *The International Journal of Quality & Reliability Management*; 2003: 20, 1; ABI/INFORM Collection, p. 20

Augustin, T. (2009). The problem of meaningfulness: Weber's law, Guilford's power law, and the near-miss-to-Weber's law, *Mathematical Social Sciences*, 57:117-130. Web. http://onlinelibrary.wiley.com/doi/10.1002/mar.4220120508/abstract

Barzilai, J. (2005). Measurement and preference function modelling, *International Transactions in Operational Research*, 12(2), 173-183, http://onlinelibrary.wiley.com/doi/10.1111/j.1475-3995.2005.00496.x/epdf

Barlett, J.E., Kotrlik, J.W., Higgins, C.C., (2001). Organizational research: Determining appropriate sample sizes in survey research, Retrieved from https://www.researchgate.net/publication/200824035_Organizational_Research_Determining_Appropriate_Sample_Size_in_Survey_Research

Britt, S.H. (1975). How Weber's law can be applied to marketing. *Business Horizons*, 18.1: 21-29. doi:10.1016/0007-6813(75)90004-X

Burke, E., Kloeber, J.M., Deckro, R.F. (2002). Using and abusing QFD scales, doi:10.1081QEN-120006707

Center for Generational Kinetics, (2016). Generational breakdown: Info about all of the generations, Retrieved from http://genhq.com/faq-info-about-generations/

Chiou, W.B. (2008). Consumers' preference shifts under the deletion of common features with varying attractiveness: An examination of competing explanations, *Psychology & Marketing*, 25: 382–398. doi:10.1002/mar.20214

Chiranjeev, K., Rajneesh, S. (2011). The price is right? Guidelines for pricing to enhance, *Profitability, Business Horizons*, doi:10.1016/j.bushor.2011.08.001

Crawford, A. (2012). Nonlinear force profile used to increase the performance of a haptic user interface for teleoperating a robotic hand, Retrieved from https://inldigitallibrary.inl.gov/sti/5517244.pdf

Decker, R., Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews, *International Journal of Research in Marketing*, http://www.sciencedirect.com/science/article/pii/S0167811610000753

Delano, G., Parnell, G.S., Smith, C., Vance, S.M. (2000). Quality function deployment and decision analysis-A R&D case study, *International Journal of Operations & Production Management*, Vol. 20 Issue: 5, p.591-609, https://doi.org/10.1108/01443570010318959

Diamond, W. (1990). Schemas determining the incentive value of sales promotions, *Psychology and Marketing*, doi:10.1002/mar.4220070303

Fechner, G.T., Boring, E.G., Howes, D.H., & Adler, H.E. (1966). Elements of psychophysics. *Translated by Helmut E. Adler, Edited by Davis H. Howes And Edwin G. Boring, With an Introduction by Edwin G. Boring*, Holt, Rinehart and Winston, Original title: *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel; 1860

Fiorenzo, F., Alessandro, P. (1999). Rating scales and prioritization in QFD, *International Journal of Quality & Reliability Management*, Vol. 16 Issue: 1, pp.85-97, https://doi.org/10.1108/02656719910250881

Franceschini, F., Rupil, A. (1999). Rating scales and prioritization in QFD, *The International Journal of Quality & Reliability Management*, Bradford Vol. 16, Iss. 1, p.85-96

Franceschini, F., Rossetto. S. (1995). The problem of comparing technical & engineering design requirements, *Research in Engineering Design*, Vol 7:270-278

Hauser, J.; Clausing, D. (1988). The house of quality, *Harvard Business Review*, vol. 66, no. 3, pp. 63–73

Hoyle, C., Chen, W., Wang, N., & Gomez-Levi, G. (2010). Understanding and modelling heterogeneity of human preferences for engineering design, Journal of Engineering Design, 22:8, 583-601, DOI: 10.1080/09544821003604496

Krider, R.E., Raghubir, P., & Krishna, A. (2001). Pizzas: π or square? Psychophysical biases in area comparisons. *Marketing Science*, 20(4), 405-425.

Krueger, L.E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law, *Behavioral and Brain Sciences*, *12*(2), p. 251-267

Kumar, M. & Noble, C. (2008). Using product design strategically to create deeper consumer
Connections, *Business Horizons*, *doi*:10.1016/j.bushor.2008.03.006

Kwong, J., Wong, K. (2006). The role of ratio differences in the framing of numerical information, *International Journal of Research in Marketing*, https://doi.org/10.1016/j.ijresmar.2006.07.001

Labrecque, L. (2013). The marketers' prismatic palette: A review of color research and future directions, *Psychology and Marketing*, doi:10.1002/mar.20597

Muunich, A. (2006). An axiomatic characterization of value judgments relative to a reference point, *Mathematical Social Sciences*, doi:10.1016/j.mathsocsci.2005.06.002

Nelson, V., Britt, S.H. (1976). The marketing importance of the "just noticeable difference", *Business Horizons*, 19: 38-40.:doi:10.1016/0007-6813(76)90063-X

Sironen, S., Leskinen, P., Kangas, A., & Hujala, T. (2014). Variation of preference inconsistency when applying ratio and interval scale pairwise comparisons, *Journal of Multi-Criteria Decision Analysis*, 183-195, Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/mcda.1500/full

Smeets, J.B.J., Brenner, E. (2016). Grasping Weber's law, *Current Biology*, Vol. 18, No. 23

Tantillo, J., Lorenzo-Aiss, J.D. & Mathisen, R.E. (1995). Quantifying perceived differences in type styles: An exploratory study, *Psychology & Marketing*, 12: 447–457, doi:10.1002/mar.4220120508

U.S. 2010 Census, (2011). 2010 Census briefs-age & sex composition, Retrieved from https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf

Vance, M. (1995). Douglas aircraft advanced common flight deck (ACF), Quality function deployment, A structured approach to selecting concepts. Unpublished corporate report, McDonnell Douglas Advanced Systems & Technology-Phantom Works Division, Systems Assessment & Planning, St. Louis, MO/The Boeing Corporation, Chicago, IL