28 August 2016

Choice Defines Value: a predictive modeling competition in health preference research

Benjamin M. Craig, PhD*
Department of Economics, University of South Florida, Tampa, Florida, USA
Mathias Barra, PhD
Health Services Research Centre, Akershus University Hospital, Lørenskog, Norway
Catharina G. M. Groothuis-Oudshoorn
Department of Health Technology and Services Research, University of Twente, Twente, The Netherlands
John D. Hartman, MA
Haas Center for Business Research and Economic Development, University of West Florida, Pensacola, Florida, USA
Elisabeth Huynh, PhD
Institute for Choice, University of South Australia, Adelaide, Australia
Michał Jakubczyk, PhD
Decision Analysis and Support Unit, SGH Warsaw School of Economics, Warsaw, Poland
Juan M. Ramos-Goñi, MSc.
EuroQol Research Foundation, Rotterdam, The Netherlands
Elly A. Stolk, PhD
EuroQol Research Foundation, Rotterdam, The Netherlands
Kim Rand-Hendriksen, PhD
Department of Health Management and Health Economics, University of Oslo, Oslo, Norway;
Health Services Research Centre, Akershus University Hospital, Lørenskog, Norway

Funding/Support: Funding support for this research was provided by the EuroQol Research Foundation and a grant from the National Institutes of Health, Department of Health and Human Services, through the National Cancer Institute (1R01CA160104). The funding agreements ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The views expressed by the authors in the publication do not necessarily reflect the views of the EuroQol Group.

Conflicts of interest: There are no conflicts of interest.
Running Title: Choice Defines Value
Word Count: 3209
Online Appendices: Screenshots, Team Submissions, Predictions
Intended Journal: Journal of Health Economics

Address correspondence to: Benjamin M. Craig, PhD, Department of Economics, University of South Florida, 4202 E. Fowler Avenue, CMC206A, Tampa, FL 33620, Phone: (813) 974-4252, Fax: (813) 974-6510, email: bcraig@usf.edu

The section in yellow will be completed after the announcement of the victorious team in September. The section in blue will be written by the victorious team leader after the announcement.

ABSTRACT

Background: Health is not bought and sold openly. Therefore, in order to understand the value of various attributes thereof, health preference researchers conduct surveys using such elicitation techniques as paired comparisons (asking, for example, “Which do you prefer?”). In the absence of external data with which to validate their findings, researchers may estimate models that are suboptimal in terms of predictive validity. Instead of conducting analyses that rely heavily on conventional practices and assumptions, Drs. Craig and Rand-Hendriksen proposed a predictive modeling competition.

Aims: To engage a large group of scientists in a modeling competition in order to identify specifications and approaches to model selection that better predict health preferences.

Methods: Announced in March 2016, the competition attracted 18 teams from around the world. In April, an exploratory survey was fielded in response to which 4074 US respondents completed 20 out of 1560 paired comparisons by choosing between two health descriptions (e.g., between longer lifespan and better health). The paired comparison data were promptly distributed to all 18 teams and were also posted online. By early July, ten teams had dropped out of the competition and eight teams had submitted their predictions for 1600 additional pairs and described their analytical approach to modeling health preferences. After these predictions had been posted online, a confirmatory study was fielded that included the 1600 additional pairs and 4148 additional respondents.

Results: In September, the victorious team, X led by X, was announced at the EuroQol Plenary in Berlin, Germany. This team achieved the smallest chi square 4391.54, and their predictions were rejected by only 134 of the 1600 pairs (8%) in the confirmatory survey at a p-value of 0.01. (MORE)

Conclusions: This crowdsourcing endeavor demonstrated the diversity of analytical approaches and highlighted the importance of predictive validity in health preference research. (MORE) Although the generalizability to specific applications is unknown, the models that are more predictive may inform and inspire debate among health preference researchers.

Crowdsourcing is the process of obtaining services, ideas, or content by soliciting contributions from a large group of people rather than by relying on a single person or a handful of collaborators. By gathering the ideas of multiple independent teams, such a communal endeavor fosters greater creativity and tends to achieve a wider range of possible solutions and perspectives. This predictive modeling competition was designed on the premise that the community of health preference researchers is diverse in modeling expertise and perspectives.(Craig et al., 2016) Instead of relying on convention, peer review, or theoretical assumptions, the competition produced a diversity of analytical approaches.

Health preference research (HPR) is a scientific enterprise: specifications are devised, hypothesized, and tested. Its mantra is "choice defines value." However, by convention HPR typically conducts just one study, estimates just one analytic specification, and promotes implementation without confirmation. It seems misguided to ground health policy decisions on preliminary evidence acquired and presented from the perspective of a single team. More troubling is the approach of researchers who estimate multiple specifications and cherry-pick their results (as in data mining).(Mannila, 1997) In clinical trials, analysis plans must be formally registered before collecting and examining the data, and the results are typically confirmed by multiple teams before putting them into practice. In addition to demonstrating a diversity of analytical approaches, this competition was designed to promote scientific rigor in HPR by having multiple teams compete in their pursuit of predictive validity.

To the best of our knowledge, this is the first predictive modeling competition in HPR. Predictive validity is particularly important in HPR, because health is not bought and sold openly. Therefore, in order to understand the value of various attributes thereof, health preference researchers conduct surveys using such elicitation techniques as paired comparisons.(Berg, 1972) For the predictive modeling competition reported here, data on paired comparisons from an exploratory survey were distributed to multiple teams so that each team might apply their own modeling specifications independently. Using their findings, each team submitted predictions for a second, confirmatory set of paired comparisons. After their predictions had been posted publicly, a confirmatory survey was fielded and the teams' submissions were ranked in accordance with their predictive validity (smallest to largest chi square). Although the competition has only one winner, this crowdsourcing endeavor was also designed to yield benefits more generally to the HPR community: by promoting greater understanding of the merits underlying alternative modeling specifications, promoting the importance of predictive validity in HPR, and demonstrating the diversity of analytical approaches among HPR researchers.

METHODS

Team registration

In March 2016, Dr. Craig and Rand-Hendriksen distributed an announcement inviting all interested teams to participate in the predictive modeling competition.(Craig and Rand-Hendriksen, 2016) By April, 18 teams had registered using a brief form on the IAHPR website (no exclusions) that asked five questions pertaining to (1) conditional agreement for teams; (2) the names of the team and team leader, and the number of co-investigators; (3) the names of the co-investigators; (4) experience; and (5) invoicing. By May, all registered teams had received the exploratory data and a sample submission. By July, eight of the 18 teams had submitted their forms and predictions, and were paid $2500 (see Appendix). In September, the victorious team received a small trophy at 2016 EuroQol Plenary in Berlin and lead authorship of this manuscript in concordance with the Vancouver criteria.(2010)

Task and Pair Selection

The design of the paired comparisons (e.g., Figure 1) was largely based on the recent protocols for the Valuation Technology developed by the EuroQol Group (EQ-VT).(Oppe et al., 2016) The wording differed from the EQ-VT in four ways. (1) Because it was designed to elicit preferences, not judgments, the competition survey instrument asked "Which do you prefer?" instead of "Which is better?" (2) The labels "A" and "B" were dropped, since they might imply rank. (3) The differentiating attributes and numbers were bolded. (4) Each description included the timing and duration of problems (e.g., "Starting today, [X] years with health problems: [health state] then die ([X] years from today)").

The set of 1560 pairs in the exploratory survey was based on the 196 pairs in the EQ-VT. Every pair had two health descriptions, each of which included five problems based on the EQ-5D-5L (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). Each problem was characterized as being at one of five possible levels (none [level 1], slight, moderate, severe, unable/extreme [level 5]). As a shorthand notation, the five problems can be characterized as a vector of five numbers (e.g., Figure 1 includes 33333). The problems based on the 196 pairs of EQ-VT (and four ancillary pairs) had durations in four different temporal units (days, weeks, months, years), creating 800 efficient pairs.

In addition to the 800 efficient pairs, 760 time trade-off (TTO) pairs of identical structure were included. In TTO pairs, however, one health description involves no health problems (i.e., 11111) and a longer lifespan (e.g., Figure 1), like a conventional TTO task. To select the TTO pairs, 38 descriptions were selected from the efficient pairs, included durations in four different temporal units, and paired with five durations with no health problems (38×4×5=760).

The set of 1600 pairs in the confirmatory survey included 800 efficient pairs and 800 TTO pairs and were created using a similar process, albeit with some important differences. We began by selecting health descriptions that commonly occur in clinical data. The motivation for emphasis on prevalent outcomes is that health preference estimates are commonly applied to summarize health outcomes captured in clinical trials as a means to better inform medical recommendations and resource allocation decisions (e.g., cost-utility analyses). After combining the problems in these prevalent descriptions with durations to create a candidate set, we applied a software package (NGENE) to select a subset set of 200 pairs by d-efficiency. (ChoiceMetrics, 2012; Norman, 2016) To select the TTO pairs, 40 descriptions were selected from the efficient pairs. Each of the efficient and TTO pairs was shown with four temporal units, creating 1600 confirmatory pairs. A full description of the process of pair selection was included in the rules of the competition and distributed in advance of team registration.(Craig and Rand-Hendriksen, 2016) All pairs are shown in the Appendix.

Health Preference Survey

Between March 21 and April 6 (exploratory) and between July 25 and August 26, 2016 (confirmatory), 8,222 US adults aged 18 and older were recruited from a nationally representative panel to participate in a 25-minute online survey. The survey instrument had four components (see Appendix): Screener, Health, Paired Comparison, and Follow-up. The Screener captured the respondent's consent and demographic and socioeconomic characteristics (Table 1). Respondents who passed the Screener (i.e., who consented and did so within demographic quotas) were asked to complete the Health Component, including a five-level question on general health, the EQ-5D-5L and visual-analogue scale (range of worst to best, 0 to 100) on general health. After viewing three examples of paired comparisons, each respondent completed 10 efficient pairs followed by 10 TTO pairs. In the Follow-up Component, respondents were asked "How would you describe this survey?" and offered eight adjectival statements (Table 2) shown in random order with three response options (Not True, Sometimes True, Often True).

Aside from their fielding dates, the only differences between the exploratory and confirmatory surveys were the pairs. The order of the pairs was randomized at the respondent level, and the two alternatives within each pair were randomized horizontally at the respondent level (i.e., left and right) such that the shorter lifespan was either always on the left or always on the right. Each of the 3200 pairs had approximately 50 responses following 18 demographic quotas (all combinations of two genders, three age groups, and three race or ethnicity groups) to promote concordance with the 2010 US Census. Screenshots of the survey instrument is provided in the Appendix.

Econometric Analysis

To aid in the interpretation of the results, the respondents are described in terms of their demographic and socioeconomic characteristics. We assessed for differential dropout and differences between the exploratory and confirmatory surveys using chi square tests. Respondents are also described in terms of their response behaviors (e.g., always choosing the left option), lexicographic patterns in the paired-comparison responses (e.g., always choosing longer lifespan), and their follow-up descriptions of the survey. Regardless of how they responded to the survey, all respondents were included in the analytical sample.

To simplify the analysis, each team received only the paired-comparison responses of the exploratory survey, not also the respondent characteristics. As part of the submission form, the teams were explicitly asked whether this exclusion made the modeling more difficult (See Appendix). Using these preference data, the teams independently estimated their models and submitted predictions for the pairs in the exploratory and confirmatory surveys, ranging from 0.000 to 1.000 (see Appendix).

To characterize fit of each team's predictions, we computed the chi square:

$$\sum N_k \times (y_k - p_k)^2 \times (y_k \times (1-y_k))^{-1}$$

In this formula, N_k is the sample size (e.g., 50 responses), p_k is the team's prediction, and y_k is the sample probability for the k^{th} pair. If a sample probability was unanimous (y_k =1 or 0), the weight, $(y_k \times (1-y_k))^{-1}$, was replaced with the Berkson weight, $(4N_k^3)/(2\times N_k-1)$.(Berkson, 1955) Although the team with the smallest chi square based on the confirmatory survey wins the competition, the chi square of the confirmatory survey (y-axis) is plotted against that of the exploratory survey as an indicator of differential fit.

To illustrate which predictions are rejected by the data at a p-value of 0.01, an immediate form of the binomial test was run for each pair and team prediction. For the victorious team, this concordance between the predictions and the confirmatory responses is shown in a scatterplot in which the red dots represent rejections (Figure 3). When the confirmatory responses on any pair reject a prediction, this suggests poor predictive validity (Table 3). We also calculated Lin's concordance between predictions and sample probabilities as an absolute measure of concordance (Lawrence, 1989) and explored which specific pairs were the most difficult to predict across all teams (as shown by the number of rejections).

To facilitate comparison of fit across pair subsets, reduced chi square (i.e., chi square divided by the number of pairs within a pair subset) was estimated by team, survey (exploratory or confirmatory), temporal unit, difference in survival (immediate death, half or less, or more than half) and pair type (efficient or TTO).(Andrae et al., 2010)

RESULTS

Among the 13,974 US adults recruited by email for the study (i.e., survey visits), 12,123 (87%) completed the screener, 9,212 (66%) were selected to participate in the study (i.e., consented and were quota sampled), 8,721 (62%) completed the health component, and 8,222 (59%) completed the paired comparisons (median 18.26 minutes; 13.52-27.02 IQR). The 990 respondents who dropped out during the Health or Paired Comparison Components were often female, Black or African American, or reported "Refused/Don't know" for household income than respondents who completed the Components (Table 1; p-value<0.01). Although there were no significant differences in respondent demographics, the exploratory and confirmatory samples have small differences in education attainment and household income (<5%). The analytical samples has higher educational attainment and household income compared to the US 2010 Census.

In Table 2, we show patterns in the response behaviors (p-value<0.01). Few respondents chose only left or only right (<0.5%). Some respondents always chose the shorter lifespan (<2%) and others (5%) always chose the alternative with the longer lifespan. These behaviors and lexicographic preferences were slightly less prevalent in the exploratory survey compared to the confirmatory survey (<3% difference). Based on the reported descriptions of the survey, most considered the survey to be "Interesting, thought provoking, eye opening" (90%) and "Challenging, tricky, tough and difficult" (78%). Less than half considered the survey "Ridiculous, implausible, unrealistic," "Enjoyable, amusing, entertaining, fun," or "Unclear, vague, and nebulous." The descriptions of the exploratory and confirmatory surveys were similarly, except that a few additional respondents (<3% difference) indicated that the confirmatory survey was "Weird, unusual, bizarre, odd, strange."

Figure 2 shows the chi square for the confirmatory and exploratory surveys by team. Among the eight teams, chi square ranged from 908.78 to 5587.42 for the exploratory survey and from 4391.54 to 8028.86 for the confirmatory survey. Among the eight teams, X led by X, submitted the predictions with the lowest chi square of the confirmatory survey (4391.54). Their predictions also had the fewest number of rejections and the highest concordance as measured by Lin's rho, suggesting that their analytical approach clearly had the greatest predictive validity.

Table 3 shows the reduced chi square by team, survey and pair types. All teams had the most difficulty predicting preferences in "years" and the least difficulty doing so in "weeks." The teams' predictive validity differed greatly for dead pairs (pairs including "immediate death") and were more similar for lifespan pairs (i.e., where a shorter lifespan was paired with a longer lifespan). All teams predicted the TTO pairs better than the efficient pairs in the exploratory survey, but this was not the case for the confirmatory survey.

Comparison of Analytical Approaches

The submission forms enabled teams to describe the process or rationale according to which they selected their model (See Appendix); and this, as we see it, may be of greater importance than the models and estimation techniques. Four of the eight teams built from the example (Fedora), which was a Bradley-Terry model with a power function to relax the constant proportionality assumption. Instead of characterizing all of the models in greater detail here, we refer the reader to the team's descriptions of their analytical approaches.

In the confirmatory survey, the four teams that used the Bradley-Terry model (Fedora, Occam's Barbershop Quartet, Preferential Treatment, and Marginal Choices) performed worse on predicting the dead pairs, but largely better on predicting the lifespan pairs compared to the other four teams, which suggests that the Bradley-Terry model (or its power function) may need to be modified for the prediction of dead pairs. Likewise, the four teams that used the Bradley-Terry model largely predicted

the efficient pairs better and the TTO pairs worse than their competitors. Based on these results, the best analytical approach may be to estimate a Bradley-Terry model to predict the efficient pairs and an alternative approach to predict the dead and TTO pairs.

(MORE ON THE VICTORIOUS MODEL)

DISCUSSION

The purpose of this study was not to conclude that a specific model should be promoted as universally "best," or as "true" in some deeper sense. The main objective was to use crowdsourcing to get as many model specifications as possible out in the open so that their strengths and weaknesses could be discussed (See Appendix). The competition successfully promoted the importance of predictive validity in HPR, and showed the diversity of analytical approaches and perspectives within the HPR community. This competition has drawn attention to critical problems with specific econometric issues and modeling approaches (e.g., logits, constant proportionality) as well as disseminated a public dataset and code so that student and scientists who are new to the field have a better understanding of the challenges of health preference modeling.

(MORE ON SUMMARY OF THE FORMS)

This project faced some challenges related to an unexpected aspect of the competition design. All 18 of the registered teams openly agreed with the amount of compensation ($2500), the rules of the competition, and the time frame for its deliverables. Chi square was selected from the set of all possible valid measures of fit as the primary means of assessing predictive validity for the competition. Knowing that they would be ranked by their chi square, each team had an incentive to submit values that minimized predictive error on the pair probabilities. Nevertheless, ten of the 18 teams dropped out after receiving the exploratory data, because (1) their intended approach performed poorly (e.g., logits); (2) the attributes involved unexpected complexities common to health valuation (e.g., different temporal units); or (3) team leaders had to attend to unexpected personal or work commitments.

When the competition was announced, some researchers expressed concern about the inherent advantages of teams involved in administering the competition. To avoid potential conflicts of interest, Dr. Craig distributed his submission (i.e., form, predictions, code) to all teams before accepting any other submissions. His submission served as an example and allowed him to review the submissions of others without provoking concern that he might then modify his own. But his example may, in turn, have contributed to the decisions of some teams to drop out and may also have induced the unintended consequence that four of the eight submissions applied a similar analytical approach to modeling, reducing analytic diversity.

Among the ten teams that dropped out, some researchers who specialize in identifying subgroups or individuals with distinct preferences (i.e., preference heterogeneity) expressed deep reservations after seeing the data, stating that predictive validity is an inherently flawed objective. From their perspective, preference data must be individually cleansed of respondent behaviors and traits before they can be properly interpreted as preferences. If one assumes that preferences are inherently latent, the prediction of confirmatory preference data is critically confounded by underlying unobservable factors. Furthermore, the selection of the chi square as the measure of fit was arbitrary (as any other measure would be) (Canary et al., 2016; Hosmer et al., 1997) and an added source of debate. The fact that ten teams dropped out after agreeing to the rules and examining the data exemplifies the diversity of analytical approaches and perspectives in HPR as well as a limitation of this competition.

Apart from its latency, we further recognize that preferences may be heterogeneous and that this competition tells us little about predicting preferences of specific individuals or differences in individual perceptions.(Craig et al., 2015) The teams predicted preferences of the general population in aggregate; however, this evidence does not imply that their models predict preferences at the individual level. Models that perform well in aggregate data may perform poorly when predicting the preferences of a specific individual, and vice versa. Individual preferences may be lexicographic (e.g., perfect complements under Leontief utility); (Leontief, 1966) and, therefore, may be poorly expressed as a continuous function. The literature on ecological fallacies includes many examples where markets act predictably but individuals do not.(Loney and Nagelkerke, 2014; Stigler, 1950) To demonstrate that a model is a suitable for individual prediction would require a respondent-specific analysis, which is planned for a future competition.

Also, it is important to acknowledge the potential biases in panel-based surveys, which is particularly challenging for experimental studies. (Craig et al., 2013) In this study, low socio-economic status (SES) is rare in online panels and associated with dropping out and with non-trading behavior (e.g., always choosing the alternative on the right or with the longest lifespan). Lexicographic response patterns may be attributable to preferences, inattentiveness, or greater cognitive difficulty. Therefore, even if online panels were able to recruit a sufficient number of participants with low SES (external validity), the responses may not reflect their actual preferences (internal validity). Furthermore, it is reasonable to expect that the exploratory and confirmatory data may differ due to seasonal or other unobservable changes in the panel. These limitations should be balanced against the feasibility of controlling such biases and its potential implication for the competition results.

(CONCLUSION)

REFERENCES

ClinicalTrials.gov[Internet]. National Library of Medicine (US): Bethesda, Maryland, USA.

Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication. Journal of Pharmacology & Pharmacotherapeutics 2010;1; 42-58.

Andrae R, Schulze-Hartung T, Melchior P. 2010. Dos and don'ts of reduced chi-squared. Cornell University Library; 2010.

Berg RL. Health Status Indexes: Proceedings of a Conference Conducted by Health Services Research, Tucson, Arizona, October 1-4, 1972s. Hospital Research and Educational Trust: Tucson, Arizona; 1972.

Berkson J. Maximum Likelihood and Minimum Chi Square Estimates of the Logistic Function. Journal of the American Statistical Association 1955;50; 130-162.

Canary JD, Blizzard L, Barry RP, Hosmer DW, Quinn SJ. Summary goodness-of-fit statistics for binary generalized linear models with noncanonical link functions. Biometrical Journal 2016;58; 674-690.

ChoiceMetrics. Ngene 1.1.1 User Manual & Reference Guides.: Australia; 2012.

Craig BM, Hays RD, Pickard AS, Cella D, Revicki DA, Reeve BB. Comparison of US panel vendors for online surveys. Journal of medical Internet research 2013;15; e260.

Craig BM, Mühlbacher A, Lancsar E, Brown DS, Ostermann J. Health Preference Research: An Overview. Medical Care 2016;Under review.

Craig BM, Pickard AS, Rand-Hendriksen K. Do health preferences contradict ordering of EQ-5D labels? Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation 2015;24; 1759-1765.

Craig BM, Rand-Hendriksen K. 2016. EQ DCE Competition Description, Rules, and Procedures v1.1. International Academy of Health Preference Research: Tampa, Florida, USA; 2016.

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine 1997;16; 965-980.

Lawrence IKL. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 1989;45; 255-268.

Leontief W. Input-output economics. Oxford University Press: New York,; 1966.

Loney T, Nagelkerke NJ. The individualistic fallacy, ecological studies and instrumental variables: a causal interpretation. Emerging Themes in Epidemiology 2014;11; 1-6.

Mannila H. 1997. Methods and Problems in Data Mining. Proceedings of the 6th International Conference on Database Theory. Springer-Verlag; 1997.

Norman R. 2016. Appendix on Ngene Pair Selection. International Academy of Health Preference Research: Tampa, Florida, USA; 2016.

Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N. EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. PharmacoEconomics 2016; 1-12.

Stigler GJ. The Development of Utility Theory. II. Journal of Political Economy 1950;58; 373-396.

Figure 1. Example of a Paired Comparison

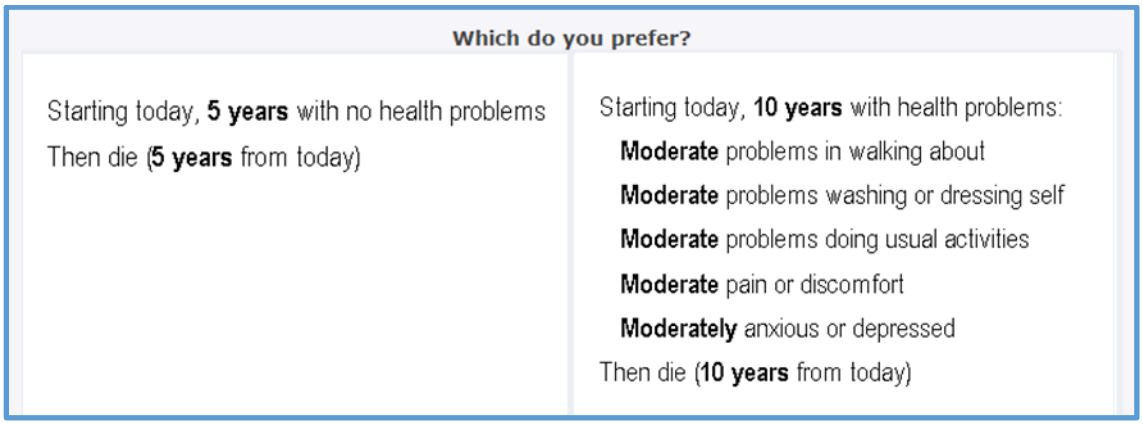
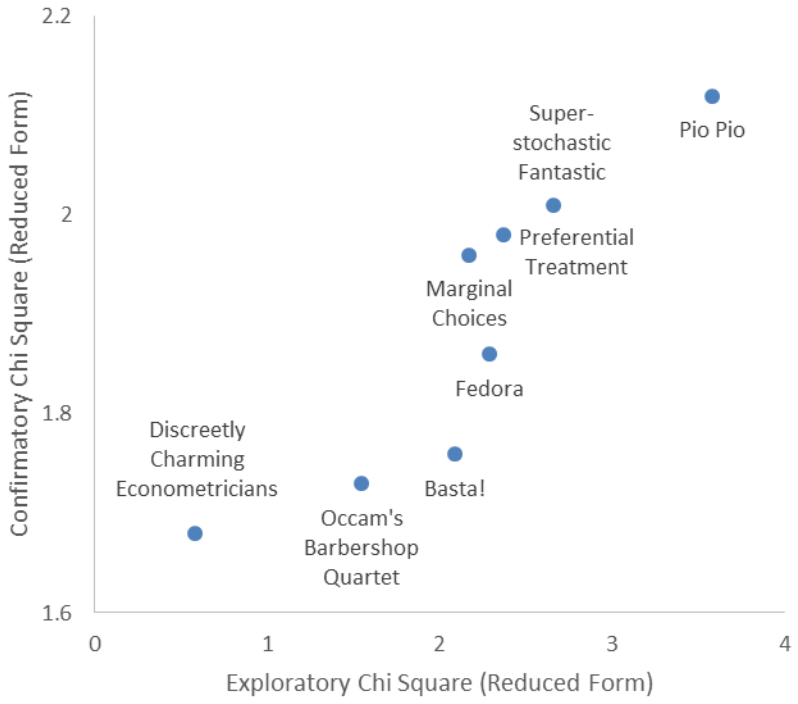
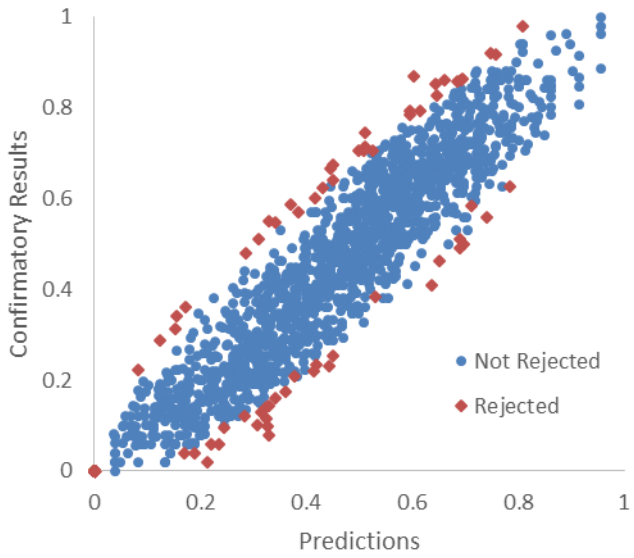


Figure 2. Exploratory and Confirmatory Chi-square (DRAFT; N=374; 10% sample only)



NOTE: This scatterplot will be updated after the announcement of victorious team in September.

Figure 3. Predictions and Confirmatory Finding: (insert victorious team)



NOTE: This scatterplot of the exploratory estimates and Fedora predictions will be replaced with the scatterplot of the confirmatory estimates and the predictions of the victorious team.

Table 1. Respondent Characteristics by Completion and Compared to 2010 US Population*

		Complete				
	Dropout N=990 % (#)	Exploratory N=4074 % (#)	Confirmatory N=4148 % (#)	Dropout versus Complete p-value	Exploratory versus Confirmatory p-value	US 2010 Census %
Age in years						
18 to 34	25.66 (254)	27.12 (1105)	28.21 (1170)	0.22	0.55	30.58
35 to 54	38.59 (382)	36.25 (1477)	35.68 (1480)			36.70
55 and older	35.76 (354)	36.62 (1492)	36.11 (1498)			32.72
Sex						
Male	42.83 (424)	49.39 (2012)	50.53 (2096)	<0.01	0.30	48.53
Female	57.17 (566)	50.61 (2062)	49.47 (2052)			51.47
Race						
White	77.58 (768)	82.18 (3348)	80.54 (3341)	0.02	0.11	74.66
Black or African American	16.36 (162)	11.73 (478)	12.73 (528)			11.97
American Indian or Alaska Native	0.51 (5)	0.56 (23)	0.48 (20)			0.87
Asian	2.83 (28)	2.82 (115)	3.01 (125)			4.87
Native Hawaiian or other Pacific Islander	0.40 (4)	0.59 (24)	0.39 (16)			0.16
Some other race	2.32 (23)	2.11 (86)	2.84 (118)			5.39
Two or more races						2.06
Hispanic ethnicity						
Hispanic or Latino	12.02 (119)	12.03 (490)	12.30 (510)	0.90	0.71	14.22
Not Hispanic or Latino	87.98 (871)	87.97 (3584)	87.70 (3638)			85.78
Educational attainment among age 25 or older						
Less than high school	2.42 (24)	1.91 (78)	2.10 (87)	0.96	<0.01	14.42
High school graduate	43.84 (434)	43.27 (1763)	45.20 (1875)			28.50
Some college, no degree	11.11 (110)	12.20 (497)	8.87 (368)			21.28
Associate's degree	6.97 (69)	7.44 (303)	5.67 (235)			7.61
Bachelor's degree	29.19 (289)	27.91 (1137)	30.30 (1257)			17.74
Graduate or professional degree	3.43 (34)	3.61 (147)	3.71 (154)			10.44
Refused/Don't know	0.10 (1)	0.07 (3)	0.07 (3)			-
Household Income						
$14,999 or less	5.35 (53)	4.52 (184)	4.44 (184)	<0.01	<0.01	13.46
$15,000 to $24,999	7.07 (70)	5.65 (230)	5.74 (238)			11.49
$25,000 to $34,999	8.59 (85)	8.27 (337)	8.20 (340)			10.76
$35,000 to $49,999	14.75 (146)	15.34 (625)	12.85 (533)			14.24
$50,000 to $74,999	20.51 (203)	21.23 (865)	21.79 (904)			18.28
$75,000 to $99,999	12.83 (127)	15.56 (634)	14.05 (583)			11.81
$100,000 to $149,999	12.63 (125)	13.21 (538)	15.41 (639)			11.82
$150,000 or more	6.67 (66)	7.51 (306)	9.81 (407)			8.14
Refused/Don't know	11.62 (115)	8.71 (355)	7.71 (320)			-

* Age, sex, race, and ethnicity estimates for the US are based on 2010 Census Summary File 1. Educational attainment and household income are based on 2010 American Community Survey 1-Year Estimates. Unlike the US Census, the American Community Survey excluded adults not in the community (e.g., institutionalized) and describes income by the proportion of households, not adults.

Table 2. Response Behavior, Lexicographic Preferences and Survey Description

	Exploratory N=4074 % (#)	Confirmatory N=3550 % (#)	p-value
Response Behavior			
Always Left or Always Right	0.34 (14)	0.48 (20)	0.33
Both	99.66 (4060)	99.52 (4128)	
Lexicographic Preference			
Always Shorter Lifespan	0.69 (28)	1.66 (69)	<0.01
Both	95.14 (3876)	91.88 (3811)	
Always Longer Lifespan	4.17 (170)	6.46 (268)	
Survey Description (ranked by frequency)			
Interesting, thought provoking, eye-opening			
Not True	10.75 (438)	9.81 (407)	0.30
Sometimes True	44.11 (1797)	44.31 (1838)	
Often True	43.47 (1771)	44.67 (1853)	
Challenging, tricky, tough, difficult			
Not True	22.21 (905)	23.41 (971)	0.51
Sometimes True	50.93 (2075)	50.36 (2089)	
Often True	25.14 (1024)	25.12 (1042)	
Weird, unusual, bizarre, odd, strange			
Not True	29.06 (1184)	27.12 (1125)	<0.01
Sometimes True	44.99 (1833)	44.17 (1832)	
Often True	23.93 (975)	27.34 (1134)	
Depressing, sad, scary, distressing			
Not True	30.83 (1256)	29.39 (1219)	0.03
Sometimes True	45.97 (1873)	45.40 (1883)	
Often True	21.33 (869)	23.79 (987)	
Morbid, morose, dismal, bleak, grim, somber			
Not True	31.81 (1296)	29.77 (1235)	0.02
Sometimes True	44.48 (1812)	44.62 (1851)	
Often True	21.89 (892)	24.37 (1011)	
Ridiculous, implausible, unrealistic			
Not True	53.17 (2166)	51.95 (2155)	0.35
Sometimes True	35.13 (1431)	36.31 (1506)	
Often True	9.89 (403)	10.51 (436)	
Enjoyable, amusing, entertaining, fun			
Not True	56.70 (2310)	55.91 (2319)	0.56
Sometimes True	31.76 (1294)	32.98 (1368)	
Often True	9.72 (396)	9.88 (410)	
Unclear, vague, nebulous			
Not True	56.77 (2313)	57.52 (2386)	0.93
Sometimes True	34.46 (1404)	34.28 (1422)	
Often True	6.92 (282)	6.92 (287)	

Table 3. Predictive validity, rejected predictions and reduced chi square by temporal unit, lifespan and pair type

	Predictive Validity		Rejected Predictions*	Reduced Chi Square**								
				Temporal Unit				Ratio of lifespans			Pair Type	
	Chi square	Lin's Rho	%	Days	Weeks	Months	Years	Immediate death	Half or less	More than half	Efficient	TTO***
Exploratory (N=4074)												
Fedora	3569.92	0.92	4.17	2.22	1.84	2.56	2.53	1.01	1.93	2.89	3.18	1.36
Occam's Barbershop Quartet	2415.13	0.94	3.40	1.57	1.44	1.53	1.66	1.03	1.43	1.75	1.78	1.32
Preferential Treatment	3704.80	0.91	6.99	2.80	1.81	1.88	3.01	5.67	2.31	2.25	2.45	2.11
Marginal Choices	3391.08	0.92	3.33	2.14	1.74	2.44	2.38	1.49	1.82	2.72	2.94	1.36
Discreetly Charming Econometricians	908.78	0.98	0.06	0.54	0.46	0.71	0.62	0.14	0.50	0.73	0.73	0.44
Pio Pio	5587.42	0.89	9.36	3.22	2.59	4.77	3.75	0.79	2.74	4.96	5.38	1.74
Basta!	3267.00	0.93	5.96	2.06	1.90	2.23	2.19	0.81	1.79	2.62	2.67	1.52
Super-stochastic fantastic	4150.17	0.92	6.22	2.60	2.33	3.18	2.54	0.58	2.24	3.40	3.88	1.42
Confirmatory (N=4148)												
Fedora												
Occam's Barbershop Quartet												
Preferential Treatment												
Marginal Choices												
Discreetly Charming Econometricians												
Pio Pio												
Basta!												
Super-stochastic fantastic												

* Rejected prediction is the proportion of pairs, where the team's prediction was rejected by the data at a p-value of 0.01 based on an immediate form of the binomial test (e.g., red dots in Figure 3).

** Reduced chi square is the chi square divided by the number of degrees of freedom (a.k.a., mean square weighted deviation). For this table, we divided by the number of pairs; therefore the reduced chi square may be interpreted as the mean of weighted squared error across the pairs.

*** The TTO pairs excludes those pairs including "immediate death," which are shown in the 8th column.

After the victorious team is announced in September, the teams will be re-order by the chi square from the confirmatory survey (column 1) and the yellow cells will be filled in.