Evaluating Respondent Burden for Maxdiff Questions

David Gleicher[1], Angela Fontes[1], Natalie Jackson[2]
Fiyin Adesina[2], Kristin Dwan[1], Eduardo Salinas[1], Rachel Miller[1]
[1]NORC at the University of Chicago, 55 East Monroe Street, Chicago, IL 60603
[2]JUST Capital, 44 East 30th Street, New York, NY 10016

Abstract
In 2018, NORC conducted cognitive interviews on two surveys for JUST Capital. A goal of these surveys was to evaluate how respondents prioritized various aspects of a company in evaluating its level of "justness." NORC conducted cognitive testing to evaluate the burden of three different question types involving prioritizing response choices.
NORC attempted to evaluate the level of burden and overall difficulty of these different types of question through two methods. The first method was a measurement of the time it took to complete the different sections comprised of the different question types. The second method was a subjective evaluation of difficulty by the respondent on a 1-5 scale for each question type after they had finished the survey.
Findings from this cognitive testing indicate very small differences in timing and subjectively assessed difficulty between different question types. Results include average subjective difficulty rating for each question time, average time to complete each question type section. Limitations and future research are also discussed.

Key Words: Maxdiff, Forced Choice, Cognitive Testing, Burden Estimation

1. Introduction

Since 2015, JUST Capital has conducted an annual survey of Americans evaluating their opinions on just company behavior[1]. This survey captures respondent opinions about the general state of American businesses as well as what they believe the priorities of companies should be. In order to capture the prioritization of aspects of business life, the survey, conducted by NORC, utilized a question type called Maxdiff. Prior to the administration of the 2018 survey, NORC conducted cognitive and usability testing of the survey instrument. This testing covered general usability, respondent understanding of terms and concepts, and burden level, comparing Maxdiff questions with two other question types that could also evaluate relative preferences. This paper will focus on the last of these three areas of cognitive testing. It will begin by describing the Maxdiff question type and will then describe the methodology used to evaluate the burden of the Maxdiff questions and the two other question types evaluated. Finally, it will conclude by discussing the results of the evaluation, limitations of the analysis, and potential avenues of future research on the topic.

---

[1] See https://justcapital.com/reports/2018-survey/ for the 2018 report.

1.1 The Maxdiff Question Type

The Maxdiff question type, also known as best-worst scaling, is a way of eliciting preferences when a respondent is forced to choose between different options. Other ways of eliciting preferences include ranking questions and questions where respondents are asked to allocate portions of a total to different options (either a percentage summing to 100 or some resource, such as money). Maxdiff is distinct from these in that it presents several different screens with different options on each, and asks the respondent to pick the most and least preferable option within each screen. This allows a large list of options to be presented in smaller groupings that are easier to evaluate for the respondent, and also enables a more fine-grained ranking of preferences[2].

2. Methodology

2.1 Background on the 2018 JUST Capital Survey

In prior years of the JUST Capital survey, respondents were asked how much of a priority companies should give to a set of categories related to running a large company. These were known as "drivers" and are listed in the table below.

Driver Categories
Workers
Customers
Products
Environment
Jobs
Communities
Leadership & Shareholders

Table 1: Drivers in JUST Capital 2018 survey

In addition, the 2018 version of the survey sought to evaluate preferences for subcategories within each of these driver categories. These were titled "components." Therefore, 4 additional surveys were cognitively tested that incorporated these "components" rather than the higher level drivers. Components from most of the drivers were combined two to a survey, except for "Leadership and Shareholders," which was in a survey by itself. These groupings and the number of subcategories within each component are listed in the table below. The number of subcategories is important because it determined the number of screens that were necessary to show the respondent in the Maxdiff section. These ranged from 3 screens of 2 options each ("environment" and "jobs" components) to 7 screens with 4 options in each ("workers" component). The fact that different components required different numbers of screens presented a challenge in terms of evaluating the burden, which will be discussed in section 2.2.2.

---

[2] Further information about Maxdiff questions can be found here:
http://www.sawtooth.com/index.php/blog/archives/maxdiff-maximum-power-to-act-on-market-preferences/

Component Grouping	Number of Subcategories for Each Component
Environment and Communities	Environment – 3 Subcategories Communities – 6 Subcategories
Products and Customers	Products – 4 Subcategories Customers – 4 Subcategories
Workers and Jobs	Workers – 9 Subcategories Jobs – 3 Subcategories
Leadership and Shareholders	Leadership and Shareholders – 7 Subcategories

Table 2: Components in JUST Capital 2018 survey, how they were grouped, and the number of response options within each component.

2.1.1 Question Type Presentation

The Maxdiff questions were presented within the same survey as two other question types evaluating the same prioritization in different ways – a ranking question and an allocation of resources question. The presentation of these three question types was randomized within each survey so that they were not presented in the same order every time and had a relatively even distribution in terms of which was presented first. The ranking and allocation questions were each presented on a single screen, whereas the Maxdiff questions necessitated multiple screens. Examples of these screens are presented in the figures below.

Figure 1: Money allocation question example

Please rank the following from 1-7, with 1 being most important to just company behavior and 7 being least important to just company behavior.

_____ PRODUCTS: A company's products and services, including fair pricing and quality, as well as the benefit or harm of the product

_____ CUSTOMERS: How a company treats its customers, including providing positive experiences, protecting their privacy, and providing fair sales terms

_____ ENVIRONMENT: A company's environmental impact, including overall environmental responsibility, using resources efficiently, and minimizing pollution

_____ WORKERS: How a company treats its employees and contractors, including fair pay, good benefits, and safe working conditions

_____ JOBS: The positive impact a company has on the job market in the U.S., including the number of people the company employs and the number of new jobs created

_____ COMMUNITIES IN THE US & ABROAD: How a company manages risks to human rights internationally, including minimizing forced labor or operating in countries with oppressive governments, and how it engages with local communities

_____ COMPANY LEADERSHIP & SHAREHOLDERS: How a company's leadership acts ethically and with integrity, achieving long-term financial growth, and creating value for its shareholders

Figure 2: Ranking question example

Which of these three categories of behaviors is most important in evaluating how just a company is, and which category is least important in evaluating how just a company is.

Most Important		Least Important
○	PRODUCTS: A company's products and services, including fair pricing and quality, as well as the benefit or harm of the product	○
○	WORKERS: How a company treats its employees and contractors, including fair pay, good benefits, and safe working conditions	○
○	CUSTOMERS: How a company treats its customers, including providing positive experiences, protecting their privacy, and providing fair sales terms	○

Figure 3: Maxdiff question example

2.2 Cognitive Interview Structure

Respondents were recruited for participation through an external panel. Cognitive interviews were conducted online through Adobe Connect. Burden was evaluated through two methods – Subjective respondent rating of burden of each question type on a 1-5 scale, and timing of how long it took the respondent to complete each question type. Because timing was being evaluated, all probing was conducted retrospectively so as not to affect this timing. All three question types were tested within each interview, and the order of presentation of each type was distributed so that it was relatively even over all interviews. The tables below show the number of interviews conducted for the drivers

and components surveys and the order of presentation of question types. Note that for the components survey, the money allocation question was changed to be a percentage allocation – distributing 100% of a company's resources for this purpose amongst the various subcategories.

Drivers (N=22)		
Maxdiff 1st	Money 1st	Ranking 1st
8	8	6

Table 3: Number of drivers interviews and distribution of the question types asked first

Components (N=41)		
Maxdiff 1st	Percentage 1st	Ranking 1st
15	11	15

Table 4: Number of components interviews and distribution of the question types asked first

2.2.2 Evaluating Burden

Because the Maxdiff question type requires presenting options several times over multiple screens, it is not straightforwardly obvious how to compare it with other question types that do not have this requirement. Comparing the objective time in seconds it takes to complete each question type is not a fair measurement. Dividing the timing by the number of screens presented is not quite right either, since the number of response options presented on each screen varies as well. The solution identified was to divide the average time per question type section by the total number of response categories the respondent saw within that section. So, for example, the average time of completion of the ranking question type for the environment component was divided by 3, since there were three subcategories presented on one page. For the Maxdiff section of the environment component, the average time was divided by 6, since there were 3 screens with 2 options each in this section. The results showing the average section timing without any division and with division by number of screens are shown in the results section for the sake of comparison, however the results using division by number of response choices provide the most accurate and comparable estimates of burden.

3. Results

Below are the tables of results for the drivers survey and components surveys, presented in order of average timing in seconds, average timing in seconds divided by the number of response choices, and subjectively self-assessed difficulty rating.
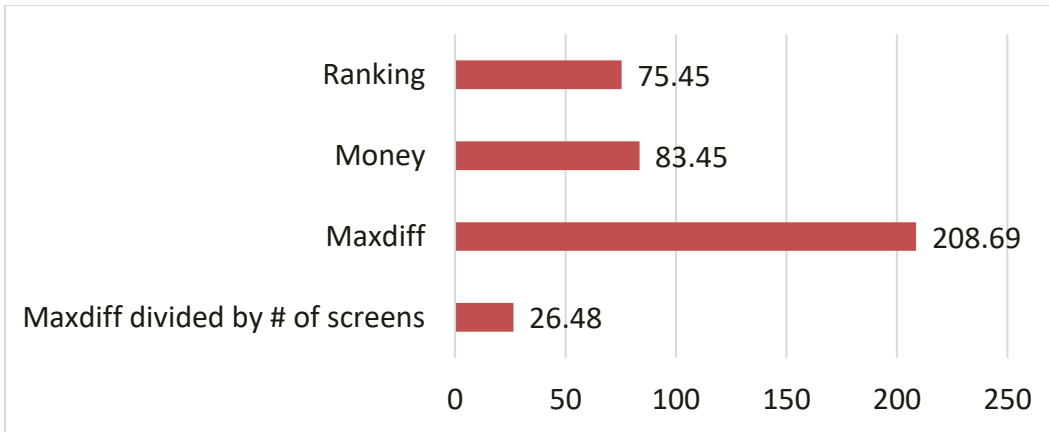
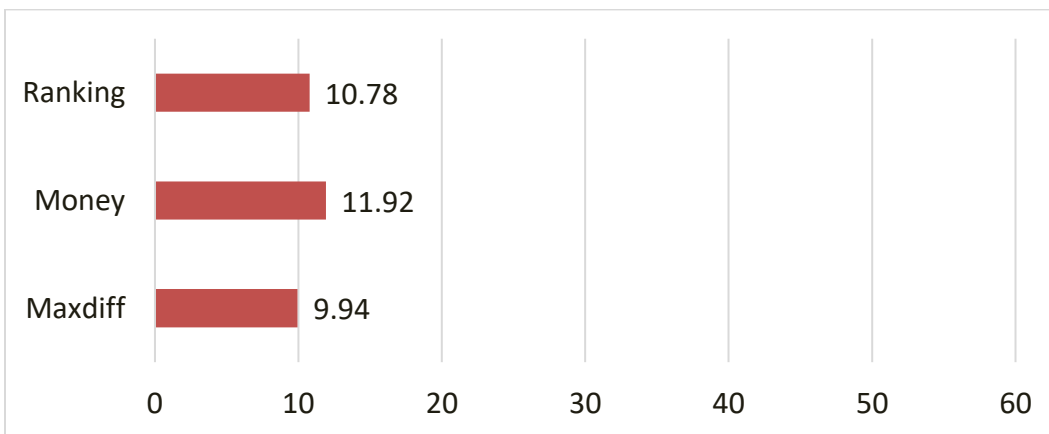**Table 5:** Drivers survey – Average section timing in seconds

Table 6: Drivers survey - Average section timing in seconds divided by number of response choices seen in that section
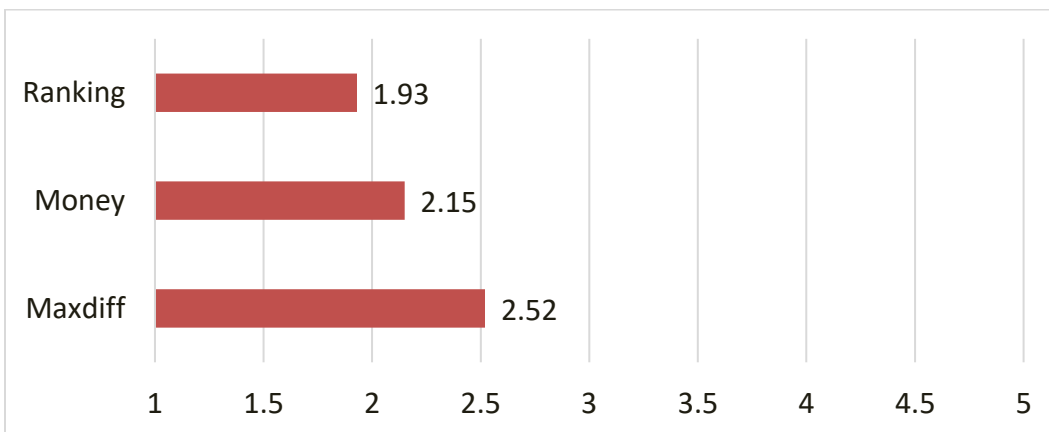

Table 7: Drivers survey - Average difficulty rating (1-5 scale, 1 last difficult, 5 most difficult)
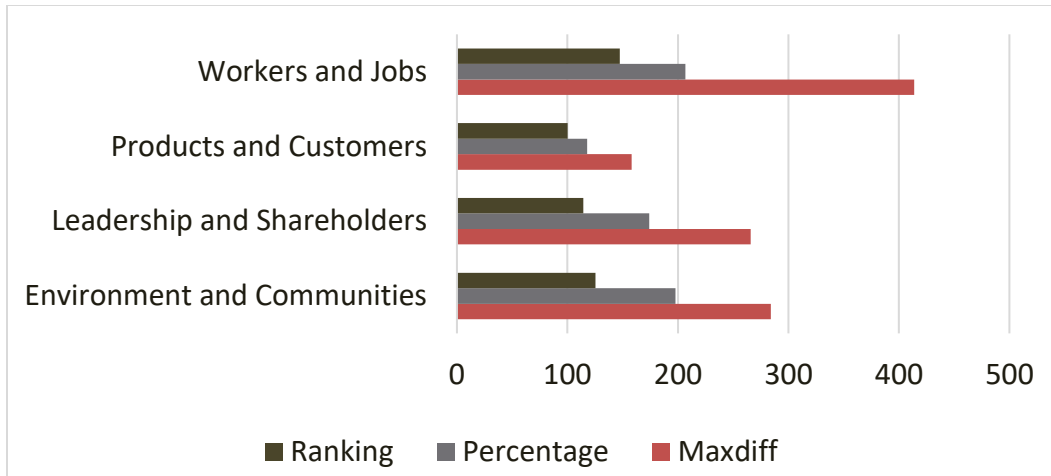
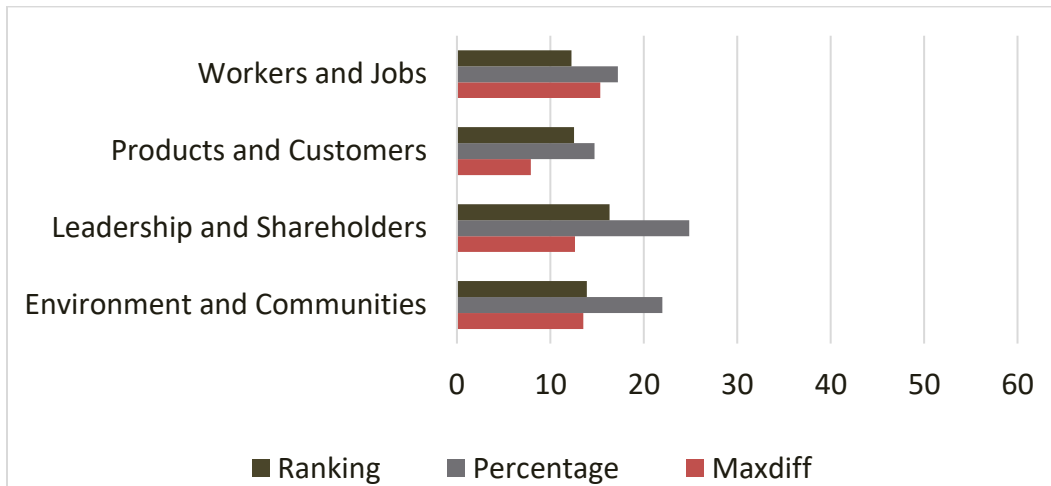**Table 8:** Components survey – Average section timing in seconds

Table 9: Components survey - Average section timing in seconds divided by number of response choices seen in that section
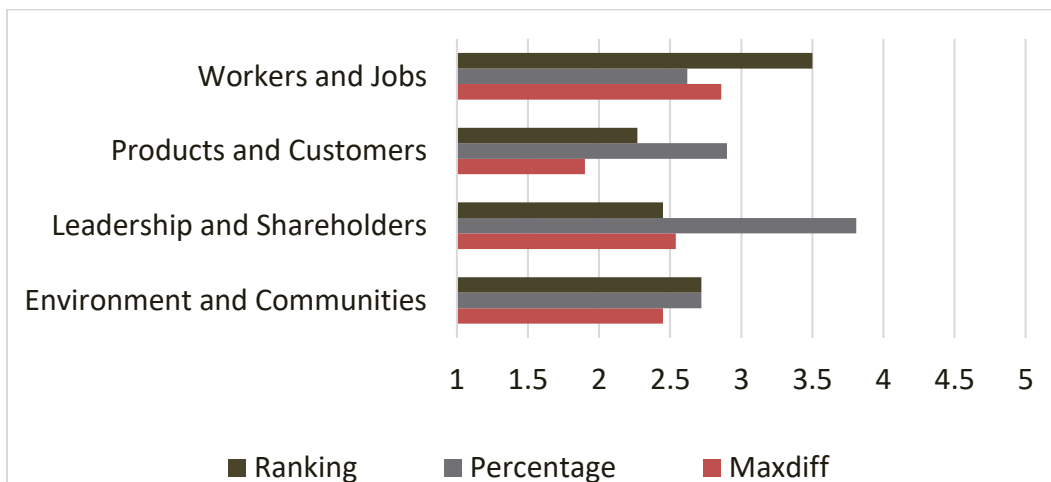

Table 10: Components survey - Average difficulty rating (1-5 scale, 1 last difficult, 5 most difficult)

3.1 Discussion

Looking at the results above, a few themes emerge. The timing differences among the different question types are extremely small once the number of response choices presented is taken into account. The components surveys were overall more difficult, both in terms of timing and subjectively rated difficulty, than the drivers survey was. Though the Maxdiff questions had the highest difficulty rating in the drivers survey, the difference was not extremely large, and this finding was not observed in the components surveys, where the Maxdiff questions did not have the highest rated difficulty in any, and had the lowest difficulty rating in 2 out of the 4 surveys.

Additionally, in retrospective probing, the most common reason reported by respondents for having difficulty was having been forced to choose between two options they felt were equally important. This provides support for using Maxdiff questions since they require a series of "most important" and "least important" evaluations rather than a strict ranking, lowering the likelihood of such conflict occurring.

Overall, the results support the idea that Maxdiff question types are not excessively burdensome compared with other question types that attempt to force a prioritization of options.

4. Limitations and Future Research

4.1 Limitations

Since this was a series of cognitive tests, the sample size was small and though the panel had demographic variation, it was not drawn to be statistically representative of the U.S. population. Although difficulty ratings were asked after the interview for each question type separately, it is not entirely clear the ratings respondents gave were fully isolated to only that question type. In order to get at this in a better way, it would be better to test each question type in its own survey that did not also have an additional question type in the same survey.

Multiple components with varying levels of difficulty were tested within the same survey. In order to better isolate the effects of question type vs. question content, a more controlled experiment that only varied question type, or held constant question type and varied a single response choice, would yield more reliable results.

4.2 Future Research

Isolating the relative burden added by additional response choices within the Maxdiff question type would be an interesting area of future research. The effects of adding additional response choices seem likely to be nonlinear in terms of additional burden, at least after a certain threshold. However, empirical investigation would be needed to confirm or refute this.

Collecting timing in surveys with a greater number of responses would yield results that have more statistical power because of the larger sample size. Along these same lines, it would be preferable to have a true experimental design where only one variable is changed between surveys while everything else remains constant (e.g., adding just one additional response choice, or one changed response choice, or only one question type per survey). Though cognitive testing is useful for identifying glaring issues or problems, there are opportunities to obtain more reliable and fine-grained measurements of the differences in question types, relative additive difficulty of response choices, or question placement, with experimental designs using a large number of completed responses.

References

Bruhn, E. 2009. MaxDiff – Maximum Power to Act on Market Preferences. http://www.sawtooth.com/index.php/blog/archives/maxdiff-maximum-power-to-act-on-market-preferences/.

JUST Capital. 2018. From Insight to Action: Just Capital's 2018 Survey Results and Roadmap for Corporate America. https://justcapital.com/wp-content/uploads/2018/10/JUSTCapital_SurveyReport_10252018.pdf