4-28-2022

Numeric Forced Rank: A Lightweight Method for Comparison and Decision-making

Erin Gannon
Google

Barbara Chaparro
Embry-Riddle Aeronautical University, chaparb1@erau.edu

Scholarly Commons Citation
Gannon, E., & Chaparro, B. (2022). Numeric Forced Rank: A Lightweight Method for Comparison and Decision-making. *CHI EA '22: CHI Conference on Human Factors in Computing Systems Extended Abstracts,* (). https://doi.org/10.1145/3491101.3503550

Numeric Forced Rank

A lightweight method for comparison and decision-making

Erin Gannon
Google
United States
ergannon@google.com

Barbara Chaparro
Embry-Riddle Aeronautical University
United States
barbara.chaparro@erau.edu

ABSTRACT

Comparing products, features, brands, or ideas relative to one another is a common goal in user experience (UX) and market research. While Likert-type scales and ordinal stack ranks are often employed as prioritization methods, they are subject to several psychometric shortcomings. We introduce the numeric forced rank, a lightweight approach that overcomes some of the limitations of standard methods and allows researchers to collect absolute ratings, relative preferences, and subjective comments using a single scale. The approach is optimal for UX and market research, but is also easily employed as a structured decision-making exercise outside of consumer research. We describe how the numeric forced rank was used to determine the name of a new Google Cloud Platform (GCP) feature, present the findings, and make recommendations for future research.

CCS CONCEPTS

• **Human-centered computing → User studies**; **Usability testing**; **Laboratory experiments**.

KEYWORDS

Rank, Comparison, Prioritization, Decision-making

ACM Reference Format:
Erin Gannon and Barbara Chaparro. 2022. Numeric Forced Rank: A lightweight method for comparison and decision-making. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3491101.3503550

1 INTRODUCTION

1.1 Background

Imagine a scenario: your company has launched a new music streaming service and needs to know which one of five genres is most important to surface on the homepage. You decide to deploy a survey to understand user preferences using a 1-5 Likert-type scale to rate importance for each genre. Results come back and three of the genres are rated roughly equally at the highly important end of the scale (i.e., ratings of "5"). Now what? Are these genres truly equally important? Comparing products, features, brands, or ideas relative to one another is a common goal in user experience (UX) and market research. Two common methods to measure relative items are: a consistent Likert-type scale across dimensions of interest like the example above (e.g., "How important is this music genre to you on a scale of 1-5?" asked individually and repeated for a set of features), and a purely ordinal stack rank (e.g., "Please stack rank this list of genres from most important to least important"), often administered by asking participants to assign each item a rank number, or by a digital drag-and-drop interface [3]. See Figure 1. Numeric rating scales can tell us about the magnitude of differences between music genres, while forced rank scales tell us about the relative positioning of items to one another without overlap.

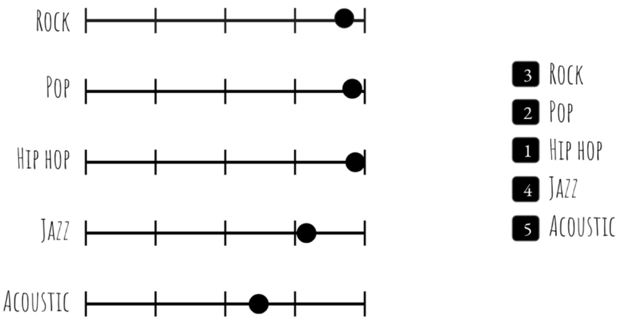

Figure 1: Traditional Likert scale (left) and stack rank (right).

Both of these methods have merits and can advance the research objectives, though they are one-dimensional in isolation; the numeric rating scales do not account for a relative comparison of the items of interest, nor do they prevent all items from being scored identically (i.e., a respondent could give each genre the same score of "5", leading to inconclusive results). Further, the stack rankings cannot tell us about the magnitude of differences. How much more do respondents prefer hip hop to classical music?

To date, few methods [16] have addressed these psychometric shortcomings in an approachable way. We introduce the numeric forced rank, a lightweight approach that can be helpful for prioritization and decision-making. The numeric forced rank overcomes some of the limitations of Likert-type scales in which respondents are often inclined to select either neutral or mid-scale responses [1, 10], presents a finer-grained scale to detect minor differences [4, 11, 12], and garners potentially greater information transmission and discriminability [8], while retaining numeric points aids in avoiding the usability issues cited in the Visual Analog Scale

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3503550

(VAS) [4, 17]. By leveraging the benefits of both numeric rating and forced rank scales, the numeric forced rank method allows researchers to collect absolute ratings, relative preferences, and subjective comments using a single scale. The approach is optimal for addressing UX and market research questions, but has also been employed as a structured decision-making exercise when teams are weighing potential alternatives for branding and product strategy.

1.2 Canonical usage

The numeric forced rank uses a long, horizontal Likert-type scale on which participants place cards representing the items (e.g., products, ideas, brands) to be ranked. Each card contains one item of interest. Items can be shown to the participant altogether, or presented one by one to probe on scoring changes as new items are introduced. Participants are not permitted to give two items the same numeric rating. A forced rank ensures that each item is given a distinct score, reduces the likelihood of ties after calculating final scores, and encourages participant to make critical judgments. The scale can be printed on paper as an in-person exercise, or the method can be completed digitally via presentation or design software, as depicted in Figure 2 and Figure 3. The numeric forced rank can be used in surveys as a purely quantitative method, but it is also used as a tool for collecting rich qualitative input alongside quantitative data. Participants are encouraged to think aloud as they contemplate card placement, and are probed on their choices as they make relative judgments.

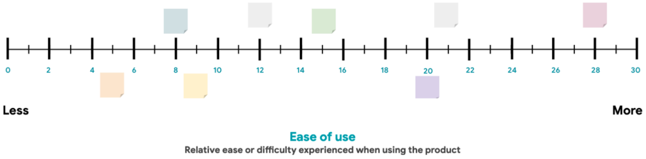
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Less
More
Ease of use
Relative ease or difficulty experienced when using the product

Figure 2: Cards representing items of interest placed at various points on the scale.

The number of scale points used will depend on the number of items to be rated and the variability expected; we typically use a minimum of three times the number of scale points as we have items to rank, rounded to the nearest 10 (e.g., a 30 pt scale for 9 items) to ensure at least 2 scale points can be placed between items if distributed uniformly, which allows space to demonstrate magnitude of differences. Scale points should be labeled from 0 to the maximum number, and the two endpoints should be labeled to indicate the left side as less and the right side as more. Scales can represent overall preference between items, a more specific prompt like ease of use, or can be used to rank products on a series of dimensions. It is often helpful to include scale labels and a definition of the dimension to be measured (see “Ease of use” in Figure 2) for the participants’ reference. The method can be used to measure a single construct, but is most powerful when employed to measure several constructs, similar to a Likert scale.

2 METHOD

2.1 Procedure

To help name a new GCP feature, our study used a 30-point scale to compare 10 potential names across five criteria our team wanted the name to meet: Descriptive, Avoids conflicts with existing industry or product terms, Scalable, Easy to read and pronounce, and Works in all interfaces in which it will appear. Criteria were determined by holding workshops on the meaning we want the name to convey, and consultation with UX/technical writers on naming best practices.

We gathered data from 8 internal participants who were familiar with the construct to be named. To accommodate a quick deadline, we ran the session as a single focus group; participants were asked to reach consensus on scores for each name per each criterion. This approach resulted in a dataset with 50 individual scores (10 names across 5 rating scales). Figure 3 shows the candidate names ranked on the “Scalable” criterion.

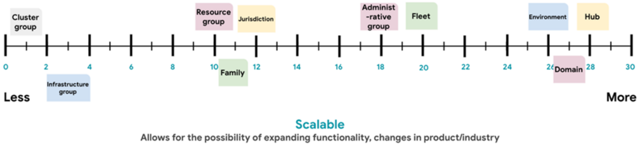

Figure 3: Name rankings for the “Scalable” criterion.

We used Google Slides on a laptop to display the scale and cards, and projected the screen to be viewed by the room. One participant was designated to move and place cards as the group discussed scores. As each name card was placed on the scale, participants were asked to elaborate on their choice and why each relative position was chosen. On the “Scalable” criterion, for example, “Cluster group” was ranked low because the feature to be named will ultimately contain more resource types than clusters.

One of the advantages of the numeric forced rank is that it is highly adaptable. This exercise can be completed in groups as described above, as an individual exercise, or as a survey to gather data from a larger sample. While multiple scales were involved in this study to represent the five criteria, a simple study may only use one scale to measure the primary dimension of interest (e.g., overall preference, ease of use, likelihood of adoption).

3 RESULTS

Given the small sample and qualitative nature of the study, we simply took the sum of scores across criteria to determine the final score of each name, and the research team recommended moving forward with the name “Hub.” See Table 1.

To understand a primary benefit of the numeric forced rank, we can examine these data as if they were purely ordinal stack rank scores (Table 2). Although “Hub” was a clear favorite and remained the top name when converting scores to a stack rank, the order of names 2-4 were changed or tied (e.g., “Cluster group” moved from position 4 to position 2). Unexpectedly, this difference became significant for our naming project; “Hub” ultimately had to be discarded due to unforeseen product changes, and we had to choose another name. Because of the more granular scoring method of

Table 1: Top 5 name scores after numeric forced rank (1=worst, 30=best).

		Criteria					
Rank	**Name**	**Descriptive**	**Does not conflict with existing industry or product terms**	**Scalable**	**Easy to read and pronounce**	**Works in all interfaces in which it will appear**	**Total**
1	Hub	3	29	28	23	28	111
2	Environment	18	6	26	14	16	80
3	Fleet	12	2	20	20	24	78
4	Cluster group	22	18	1	16	12	69
5	Admin group	21	12	18	5	6	62

Table 2: Top 5 names as stack rank scores (1=worst, 5=best).

		Criteria					
Rank	**Name**	**Descriptive**	**Does not conflict with existing industry or product terms**	**Scalable**	**Easy to read and pronounce**	**Works in all interfaces in which it will appear**	**Total**
1	Hub	1	5	5	5	5	21
2	Cluster group	5	4	1	3	2	15
3	Environment	3	2	4	2	3	14
4	Fleet	2	1	3	4	4	14
5	Admin group	4	3	2	1	1	11

the numeric forced rank, we were able to move forward with our second-highest name: "Environment." Had we taken stack rankings instead, our second choice would have been "Cluster group," which scored especially poorly on the "Scalable" and "Works in all interfaces" criteria. The magnitude of differences across criteria would have been lost in the coarse nature of stack rankings.

More rigorous analysis can be completed with a larger dataset. The proper statistical treatment of Likert scale data is a subject of contention; while some recommend treating Likert data as ordinal and applying nonparametric statistics [2], research has shown that increasing the number of scale points can approximate interval data, and that parametric tests can be performed provided test assumptions are met [14, 18]. Considering these findings, we believe the long scale format and use of multiple scales in the numeric forced rank justifies the use of parametric tests for large samples, given assumptions of group independence, normality, and homogeneity of variance are met.

4 DISCUSSION

The numeric forced rank was instrumental in choosing a name for our new GCP feature. This naming decision had been a challenge across multiple product teams for over two years. Given the difficulty that preceded our study, product stakeholders appreciated the rigor and structure the method affords. The qualitative data we collected helped provide a rationale for the decision to the broader organization, and that data continues to inform our product strategy and documentation.

In particular, the method allowed us to take advantage of the psychometric benefits of traditional Likert-type scales, while capitalizing on principles of mathematical psychology; semi-order psychological measurement models show that ranking multiple items against one another enables differentiation from usually insensitive data [5]. Simply put, rankings with 3+ items encourage participants to discriminate more critically between multiple options, especially in cases where two options appear approximately similar to each other. This echoes findings in behavioral science and judgment/decision-making literature suggesting that people are better at making comparative than absolute judgments [9, 13].

Further, the greater number of scale points affords more sensitive judgments, which can impact how items are ranked when new options are introduced: if item A is ranked at 10, Item B is ranked at 20, and Item C is ordinally between them, a respondent must decide how near or far Item C is from Items A and B. In doing so, they may decide Item B is actually even further from item A and re-score it to 25. The numeric forced rank allows us to capture those sensitivities where they would otherwise be obscured by purely ordinal scales.

It should be noted that the numeric forced rank is less structured and robust than more mathematically sophisticated approaches like Maxdiff [7]. Methods like Maxdiff and HL*m* [15] may also be better suited for lists of over 10 items [6]. The numeric forced rank was designed to be a lightweight alternative to heavier best-worst scaling techniques, and should not replace these methods when research questions and resources render them more suitable. It prioritizes simplicity and ease of entry to researchers who need a quick solution that is easy to administer and easy for participants to grasp. It also has some theoretical advantages over best-worst scaling methods in that participants are able to adjust scores relative to the entire item set through the duration of the study. They're also able to see and evaluate their final outcome to ensure it reflects their true sentiment, and adjust accordingly. Instead, Maxdiff prompts participants to repeatedly select the "best" and "worst" items from randomized subsets of the exhaustive list.

5 LIMITATIONS

The numeric forced rank has some limitations that should be considered. From a practical perspective, the exercise can be time-consuming and tedious for participants with a high number of items or scale dimensions. To mitigate, we recommend limiting the number of items to be ranked as much as possible (target 10 or fewer per scale) and being considerate of how often the method is employed. Rather than letting it replace every stack rank item in a study, the method is best when teasing apart homogenous or difficult-to-compare alternatives.

Research is limited on the reliability and validity of ranking multiple items on a single scale, though Sung and Wu [16] employed a similar approach in which the method reduced response-style bias and leniency bias when compared with Likert-type scales. Future research should focus on evaluating the test-retest reliability of ranking multiple items on a single scale, and comparing the performance of the numeric forced rank to traditional Likert-type items and stack ranks. Research should also explore the maximum number of scale items and number of items to be ranked before returns begin diminishing.

Though the numeric forced rank is ideally administered to individuals, our study employed a focus group due to limited time and resources. As such, the limitations of focus groups apply here; namely, participants may be biased by each other's answers, and/or the conversation and ranking scores may be dominated by one or a few more vocal participants. To help remedy this possibility, we asked each participant to state their preferred ranks and rationale for each criterion, starting with a different randomly selected participant each time.

6 CONCLUSION

The numeric forced rank draws from established mathematical and behavioral science theory to provide an approachable way for practitioners to gather comparison data alongside qualitative rationales. The long-scale, multi-item technique leverages the advantages of traditional scales while overcoming common shortcomings that can lead to inconclusive or conflicting results [1, 10]. To date, the numeric forced rank has been used by several teams across Google to make systematic product decisions and add rigor to otherwise unstructured processes. Beyond research studies, the numeric forced rank has also been applied among teams for structured discussion and comparisons. We hope it can continue to be a valuable tool in UX research practice and organizational decision-making.

REFERENCES

[1] Gerald Albaum. 1997. The Likert scale revisited. *Market Research Society. Journal.* 39, 2 (1997), 1–21.

[2] Phillip A Bishop and Robert L Herron. 2015. Use and misuse of the Likert item responses and other ordinal measures. *International journal of exercise science* 8, 3 (2015), 297.

[3] Jörg Blasius. 2012. Comparing ranking techniques in web surveys. *Field Methods* 24, 4 (2012), 382–398.

[4] Michelle Briggs and José S Closs. 1999. A descriptive study of the use of visual analogue scales and verbal rating scales for the assessment of postoperative pain in orthopedic patients. *Journal of pain and symptom management* 18, 6 (1999), 438–446.

[5] Clyde Hamilton Coombs, Robyn M Dawes, and Amos Tversky. 1970. Mathematical psychology: An elementary introduction. (1970).

[6] Donald R Cooper, Pamela S Schindler, and Jianmin Sun. 2006. *Business research methods*. Vol. 9. Mcgraw-hill New York.

[7] Adam Finn and Jordan J Louviere. 1992. Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing* 11, 2 (1992), 12–25.

[8] Wendell R Garner. 1960. Rating scales, discriminability, and information transmission. *Psychological review* 67, 6 (1960), 343.

[9] Richard D Goffin and James M Olson. 2011. Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science* 6, 1 (2011), 48–60.

[10] Eric A Greenleaf. 1992. Measuring extreme response style. *Public Opinion Quarterly* 56, 3 (1992), 328–351.

[11] CRB Joyce, DW Zutshi, V Hrubes, and RM Mason. 1975. Comparison of fixed interval and visual analogue scales for rating chronic pain. *European journal of clinical pharmacology* 8, 6 (1975), 415–420.

[12] Ludger Klimek, Karl-Christian Bergmann, Tilo Biedermann, Jean Bousquet, Peter Hellings, Kirsten Jung, Hans Merk, Heidi Olze, Wolfgang Schlenter, Philippe Stock, et al. 2017. Visual analogue scales (VAS): Measuring instruments for the documentation of symptoms and therapy monitoring in cases of allergic rhinitis in everyday health care. *Allergo journal international* 26, 1 (2017), 16–24.

[13] Jum C Nunnally. 1976. Psychometric theory. (1976).

[14] Mariah L. Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C. Gombolay. 2020. Four Years in Review: Statistical Practices of Likert Scales in Human-Robot Interaction Studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/3371382.3380739

[15] Jolene D Smyth, Kristen Olson, and Allison Burke. 2018. Comparing survey ranking question formats in mail surveys. *International Journal of Market Research* 60, 5 (2018), 502–516.

[16] Yao-Ting Sung and Jeng-Shin Wu. 2018. The visual analogue scale for rating, ranking and paired-comparison (VAS-RRP): a new technique for psychological measurement. *Behavior research methods* 50, 4 (2018), 1694–1715.

[17] Amelia Williamson and Barbara Hoggart. 2005. Pain: a review of three commonly used pain rating scales. *Journal of clinical nursing* 14, 7 (2005), 798–804.

[18] Huiping Wu and Shing-On Leung. 2017. Can Likert scales be treated as interval scales?—A Simulation study. *Journal of Social Service Research* 43, 4 (2017), 527–532.