Multidimensional Rating Systems: An Outlook and Exploration of Possibilities

Christian Brinkhaus[1,a)] Andrew Vargo[1,b)] Koichi Kise[1,c)]

Abstract: Expressing opinions on product satisfaction, research surveys, and market analysis has become an everyday task in our society. These statistical analyses provide important information on market research and evaluation tools. However, traditional rating systems such as the Likert scale often fail to capture nuanced feedback, as they are limited to binary or ordinal selection. The lack of adaptability can lead to survey fatigue and introduce unwanted agreement or cultural bias into the analysis. The ordinal nature of the Likert scale makes it especially difficult to use on mobile devices. We previously introduced the modified four-quadrant gradient rating scale (mFQS), a visual rating system that enables simultaneous evaluation of two dimensions through spatial interaction. Our comparative study indicated that the mFQS produces broader rating distributions and reduces cultural bias among Japanese participants compared to Likert scales. This work explores possible applications for multidimensional rating scales to broaden the landscape of rating systems in human-computer interaction research.

1. Introduction

Rating systems are a prevalent part of modern digital life. From product reviews on Amazon [*1] to user experience surveys, people constantly evaluate their experiences through surveys and questionnaires. These tools collect information ranging from qualitative observations [12], [17] to opinions [2], [13] for both companies and researchers. However, predominant paradigms such as Likert scales and star ratings are limited to ordinal or binary selection. This simplification, while convenient, introduces systematic problems: cultural bias [14], [18], [26], [27], response fatigue [3], agreement bias [22], longer response time and increased cognitive load on mobile devices [8], [16], and loss of contextual relationships between dimensions. In previous work, we have introduced the modified four-quadrant gradient rating scale (mFQS) [5] as a multidimensional rating tool that visually connects two factors in one rating (Figure 1). Although current systems are straightforward, this work opens the discussion on combining multidimensional rating scales with existing approaches to provide greater flexibility for survey evaluations.

Current research has identified several limitations in traditional rating systems. Cultural factors significantly influence rating behavior [14], [18], [26], [27], with Asian populations showing a marked midpoint tendency on Likert scales [14], [18]. The repetitiveness of long Likert scale questionnaires induces response fatigue, leading to straight-lining and reduced data quality [3]. Agreement bias causes respondents to favor positive ratings regardless of content [22]. Additionally, the limited screen space of mobile devices creates challenges for traditional surveys [6]. Only a limited body of literature explores alternative rating paradigms, such as one-dimensional sliders [7], [11]. Existing approaches such as the Affect Grid [24] demonstrate feasibility, but remain underexplored and limited to specific domains.

We previously introduced mFQS [5] (Fig. 1), a multidimensional rating interface that addresses these limitations through visual-spatial design at ACM Ubicomp 2025. Our initial evaluation in the music domain shows that the mFQS produces significantly broader distributions and possibly reduces cultural bias among Japanese participants compared to standard Likert scales. The interaction time did not show significant differences, indicating a similar cognitive load despite the novel interface.

Based on these previous results, we want to open the discussion of adapting multidimensional scales in a broader context. They are intended to complement current systems, not replace them. We see potential in using multidimensional systems alongside gold standards such as the Likert scale in questionnaires. Highly subjective and personalized domains can be difficult to evaluate with one-dimensional tools alone. Additionally, multidimensional scales offer an effective design for mobile devices with limited screen space by utilizing screen space more efficiently. Continuous selection options also facilitate easier interaction on mobile devices.

This work opens the design space for multidimensional rating systems. We identify three key opportunities:

(1) Enabling richer assessment in domains requiring nuanced evaluation factors
(2) Addressing systematic biases in traditional rating systems
(3) Adapting multidimensional ratings for mobile form factors

We want to open up the discussion on using multidimensional rat-

[1] Osaka Metropolitan University
[a)] sw25858f@st.omu.ac.jp
[b)] awv@omu.ac.jp
[c)] kise@omu.ac.jp
[*1] https://www.amazon.co.jp/

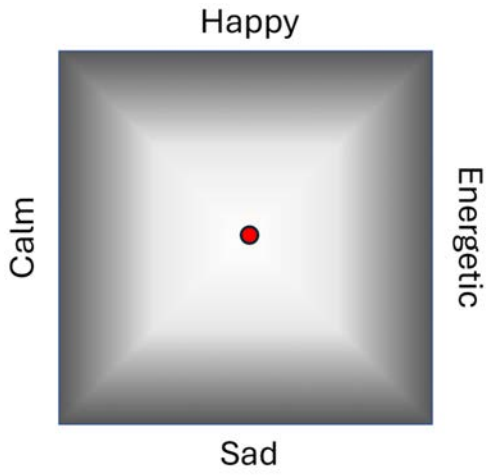

Fig. 1 The modified four-quadrant scale (mFQS) with labeled with ”calm,” ”energetic,” ”sad,” and ”happy”.

ing systems in combination with current systems and summarize potential research opportunities.

2. Background

In our previous work, we introduced the mFQS [5]. As an initial exploration, we conducted an experiment with 20 participants (10 Japanese and 10 of seven other nationalities) who rated their emotional responses to music. The music clips were evaluated on two Likert scales: calm to energetic and sad to happy. Similarly, the mFQS x-axis was labeled calm to energetic, whereas the y-axis was labeled sad to happy (Fig. 1). The participants evaluated 80 music clips in total: 20 of which had a pair and 40 others that were introduced as noise. A pair was defined as two music clips from the same song with a set offset. A pair was evaluated on both scales for direct comparison.

Analysis showed that the rating distribution of the mFQS' x-axis was statistically significantly broader than on the respective Likert scale. In addition, this was especially apparent for the Japanese sub-group of the participants, leading to indication of reduced cultural bias. Comparing rating times between scales revealed no significant differences, indicating a similar cognitive load while rating.

3. Opening the Design Space

In our first exploration, we focused on music as a domain and a two-dimensional approach in a desktop environment [5]. Our first analysis shows promising results for further exploration. We currently depend on scales that are known to introduce bias in various ways [3], [14], [18], [22], [26], [27], are not fit for mobile applications with limited screen space [6], [8], [16], and restrict users to overly coarse binary or ordinal selection. In this section, we discuss possible solutions to these problems using multidimensional scales in combination with current systems. In addition, we introduce new design ideas for such scales allowing for customization into different domains.

3.1 New Design Forms

Typical evaluation systems are limited to one-dimensional rating systems, which can reduce contextual relationships between correlated factors. Domains that need to analyze highly subjective and nuanced questions are stuck with systems that fail to connect important dimensions and only offer a limited number of selections.

Our first exploration of multidimensional scales is limited to a square two-dimensional layout (Fig. 1). This approach can be extended to additional form factors such as radial or triangular layouts. Extensions into a three-dimensional space are feasible, for example, by visualizing the scale as a cube. Even higher dimensions could be used; however, these would require combinations of multiple scales for feasible visualization.

In the following, we explore the extension into two possible domains:

Healthcare Pain assessment is essential before treatment, but pain is multidimensional. Current practice uses multiple one-dimensional rating scales and questionnaires [4]. A multidimensional tool could improve ratings by combining intensity (mild to severe) and pain type (dull to sharp) into a single linked scale. Similar approaches could be used for depression, where existing tools rely on numerical scales and questionnaires [25] but ignore multidimensionality. Integrating multiple questions into one multidimensional scale can yield better scores, higher accuracy, and clearer relationships between features.

Political Discourse The widespread use of social media for information sharing has created a new issue in news recommendation: algorithms tend to show users only content they like, forming a ”social bubble” that narrows perspectives and reinforces bias [20]. As Vargo and Tag [28] suggest, systems such as mFQS [5] can rate news articles by political stance. The multidimensional design shows correlations between the questions and their visual links, enabling intuitive rating and serving as a feedback system. Only one interaction is required to rate two connected factors.

3.2 Combination with Current Systems

Long Likert scale questionnaires can result in response fatigue [3] and introduce unwanted bias [22] over time by repetition of the same scale type.

Response fatigue occurs when participants become tired or mentally exhausted while answering a questionnaire, leading to decreased reliability or response frequency [3]. Current research suggests that limiting the number of responses or breaking the question down to a forced binary question style keeps this effect minimal [10]. However, this limits the fidelity of the responses which, for certain domains, might not be a satisfactory option.

Agreement bias [22] causes survey participants to increasingly agree with statements over time, often producing contradictory results. This can be resolved by carefully constructing Likert scale questionnaires with negated items [29], but it places a lot of responsibility on the researcher designing the questions.

Similarly, the Likert scale tends to introduce bias based on nationality [14], [18], [26], [27], which can lead to different results during analysis, possibly leading to misinterpretation. This effect is especially apparent in the Asian population, which tends to rate towards the middle of the scale [14], [18]. Due to this bias,

a straightforward interpretation of ratings is often unreliable.

Multidimensional scales in various forms could be used to reduce these effects. The intention should not be to replace current systems but to use our approach in combination with them. Response fatigue and agreement bias could be reduced by using the Likert scale and other multidimensional scale designs in combination during a questionnaire. By introducing a variety of different rating systems, we see a high possibility of giving survey participants a more varied experience while answering. Furthermore, our first analysis showed that Japanese participants rated more towards the extreme ends of the mFQS, leading to a possible alleviation of cultural bias [5], and opening opportunities for different design paradigms.

3.3 Mobile Form Factor

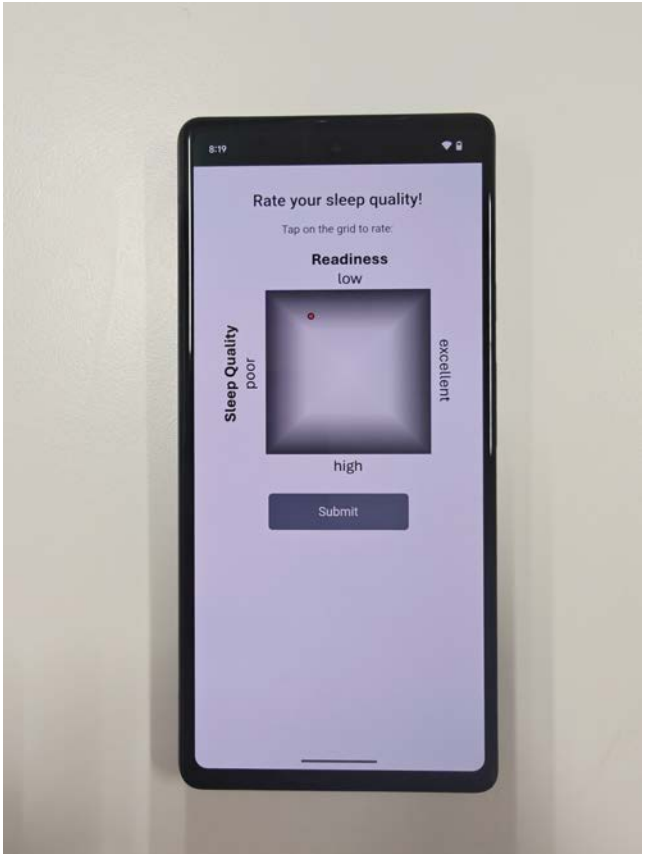

Fig. 2 The modified four-quadrant scale (mFQS) on a Google Pixel 6a. The scale is labeled with with ”poor,” ”excellent,” ”low,” and ”high” in the context of self assessing sleep quality.

With the widespread availability of mobile devices such as smartphones and smartwatches, researchers must consider the form factor of the device when designing a study. Compared to a laptop or desktop PC, a smartphone has very limited screen space. The limited screen size makes it difficult to select and interpret multiple Likert scales, as scrolling may be required to view all options [6]. In addition, the response time might be longer compared to a desktop environment [8], [16] due to misclick and increased cognitive load.

We argue that multidimensional scales are a competitive design choice for mobile devices as the screen is used more effectively (Fig. 2). The multidimensional square design fills the screen space more effectively and makes it easier to select a rating. The granularity of the scale leads to continuous point selection, and correcting an answer is easier as there are no discrete points.

Multidimensional rating systems have the potential to also be used on smartwatches. Ponnada et al. [23] showed that having an ecological momentary assessment (EMA) on a smartwatch produces a higher response rate compared to a smartphone. The responses provided on the smartwatches were shown to be less burdensome, although the frequency was much higher [23]. However, selection was limited to “Yes/Sort of/No” questions.

It is difficult to fit a Likert scale on the limited scale of the smartwatch screen. By adapting multidimensional scales to the screen size and form of a smartwatch, they can be deployed on very limited form factors. A radial layout could be used for round screens, whereas a square design is better applicable for square screen dimensions. With additional interaction design, such as first selecting the quadrant and then selecting a rating, they could open up possibilities for more nuanced in-the-moment evaluations than using a smartphone.

4. Future Work and Outlook

4.1 Comprehensive Validation

Although our initial study provides promising results in the music domain and a square two-dimensional design [5], comprehensive validation across diverse domains and contexts is essential. In addition, each domain presents unique requirements for the axis labels and differences in the rating behavior. In future studies, we want to further test reliability, validity and sensitivity [9]. We also want to explore different form factors, such as radial or triangle layout and higher dimensionality to see how these design factors affect ratings.

In our current research, we are deploying the mFQS [5] in an in-the-wild study on mobile devices to self-evaluate sleep. Although wearable devices measure objective sleep metrics [1], [15], subjective perception often differs from device-generated scores [21]. The mFQS enables users to simultaneously rate sleep quality and readiness, capturing the multidimensional nature of the sleep experience. Previous work shows that combining subjective ratings with objective measurements provides stronger predictors of cognitive performance than either alone [19]. However, traditional ordinal scales make self-assessment a burdensome task. In contrast, the visual interface of the mFQS allows users to naturally express relationships between correlated dimensions through a single intuitive interaction (Fig. 2).

4.2 Open-Source Toolkit

To enable broader adoption, we are developing an open-source toolkit for mobile and desktop devices. With this step, we want to open up the framework for more researchers who want to explore multidimensional rating systems in their study. Providing customizability of axis labels and granularity, multiple visualization styles, responsive design for desktop and mobile, and different shapes, the toolkit will allow researchers to rapidly prototype multidimensional scales tailored to their specific domains without requiring custom development.

4.3 Form Factor Adaptation

Different devices present different design challenges. Desktop interfaces benefit from precise mouse control and large screens. Mobile phones require touch-optimized interactions with limited screen space, but are more portable than a desktop PC or laptop. Smartwatches present extreme size constraints, but offer momentary assessment opportunities [23]. In the future, we will explore adaptive designs of multidimensional rating interfaces that can be used across form factors more consistently.

5. Conclusion

Traditional rating systems impose constraints such as binary or ordinal selections on aspects that are much more complex. They are plagued by bias [14], [18], [22], [26], [27], response fatigue [3], and lead to higher cognitive load on mobile devices [6]. Our mFQS [5] serves as an example of multidimensional rating systems that may address these limitations. Furthermore, we aim to expand the discussion to alternative form factors such as radial or triangular scales and extensions to higher dimensions including 3D. The goal should not be to replace current systems but to use both alongside each other, fostering the strengths of both. Breaking up the repetitiveness of the Likert scale survey with a mix of both could already help to lessen bias. However, we acknowledge that our scale needs more and deeper evaluation.

This work opens the discussion on how to utilize multidimensional rating systems more in research. We identified three key opportunities: domain-specific applications that require nuanced feedback, possible bias reduction by introducing new interface designs into research, and mobile adaptation for ubiquitous assessment. We discuss how multidimensional scales can be used in combination with current systems to achieve better capabilities for survey evaluation.

Future work will focus on comprehensive validation across domains, development of an open-source toolkit that enables broader adoption, and exploration of form factor adaptations like smartphones and smartwatches. Currently, research evaluation tools are limited to simple ordinal or binary scales. With our work, we want to open up the discussion on the use of novel multidimensional scales in combination with current systems to improve the data collection of topics that are highly subjective.

Acknowledgments This work was supported in part by the JSPS Fund for the Promotion of Joint International Research (International Collaborative Research) under Grant No. 23KK0188.

References

[1] Altini, M. and Kinnunen, H.: The Promise of Sleep: A Multi-Sensor Approach for Accurate Sleep Stage Detection Using the Oura Ring, *Sensors*, Vol. 21, No. 13, p. 4302 (online), DOI: 10.3390/s21134302 (2021). Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Berinsky, A. J.: Measuring Public Opinion with Surveys, *Annual Review of Political Science*, Vol. 20, No. Volume 20, 2017, pp. 309–329 (online), DOI: https://doi.org/10.1146/annurev-polisci-101513-113724 (2017).

[3] Birkett, N. J.: Selecting the number of response categories for a Likert-type scale, *Proceedings of the American statistical association*, Vol. 1, No. 1, pp. 488–492 (1986).

[4] Breivik, H., Borchgrevink, P.-C., Allen, S.-M., Rosseland, L.-A., Romundstad, L., Breivik Hals, E., Kvarstein, G. and Stubhaug, A.: Assessment of pain, *British journal of anaesthesia*, Vol. 101, No. 1, pp. 17–24 (2008).

[5] Brinkhaus, C., Vargo, A., Tag, B., Großmann, N., Dengel, A. and Kise, K.: (Forthcoming) mFQS: A Multidimensional Visual Rating System, *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp Companion '25, New York, NY, USA, ACM, pp. 1–5 (online), DOI: 10.1145/3714394.3754421 (2025).

[6] Buskirk, T. D. and Andres, C.: Smart surveys for smart phones: Exploring various approaches for conducting online mobile surveys via smartphones, *Survey Practice*, Vol. 5, No. 1 (2012).

[7] Cook, C., Heath, F., Thompson, R. L. and Thompson, B.: Score Reliability in Webor Internet-Based Surveys: Unnumbered Graphic Rating Scales versus Likert-Type Scales, *Educational and Psychological Measurement*, Vol. 61, No. 4, p. 697–706 (online), DOI: 10.1177/00131640121971356 (2001).

[8] Couper, M. P. and Peterson, G. J.: Why Do Web Surveys Take Longer on Smartphones?, *Soc. Sci. Comput. Rev.*, Vol. 35, No. 3, p. 357–377 (online), DOI: 10.1177/0894439316629932 (2017).

[9] Cummins, R. A. and Gullone, E.: Why we should not use 5-point Likert scales: The case for subjective quality of life measurement, *Proceedings, second international conference on quality of life in cities*, Vol. 74, pp. 74–93 (2000). Issue: 2.

[10] Dolnicar, S., Grün, B. and Leisch, F.: Quick, Simple and Reliable: Forced Binary Survey Questions, *International Journal of Market Research*, Vol. 53, No. 2, pp. 231–252 (online), DOI: 10.2501/IJMR-53-2-231-252 (2011).

[11] Funke, F., Reips, U.-D. and Thomas, R. K.: Sliders for the Smart: Type of Rating Scale on the Web Interacts With Educational Level, *Social Science Computer Review*, Vol. 29, No. 2, p. 221–231 (online), DOI: 10.1177/0894439310376896 (2010).

[12] Gerring, J.: Qualitative Methods, *Annual Review of Political Science*, Vol. 20, No. Volume 20, 2017, pp. 15–36 (online), DOI: https://doi.org/10.1146/annurev-polisci-092415-024158 (2017).

[13] Kitchenham, B. A. and Pfleeger, S. L.: Personal Opinion Surveys, *Guide to Advanced Empirical Software Engineering* (Shull, F., Singer, J. and Sjøberg, D. I. K., eds.), Springer London, London, pp. 63–92 (online), DOI: 10.1007/978-1-84800-044-5_3 (2008).

[14] Lee, J. W., Jones, P. S., Mineyama, Y. and Zhang, X. E.: Cultural differences in responses to a likert scale, *Research in Nursing & Health*, Vol. 25, No. 4, pp. 295–306 (online), DOI: 10.1002/nur.10041 (2002). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nur.10041.

[15] Malakhatka, E., Al Rahis, A., Osman, O. and Lundqvist, P.: Monitoring and Predicting Occupant's Sleep Quality by Using Wearable Device OURA Ring and Smart Building Sensors Data (Living Laboratory Case Study), *Buildings*, Vol. 11, No. 10, p. 459 (online), DOI: 10.3390/buildings11100459 (2021). Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[16] Maslovskaya, O., Smith, P. W. and Durrant, G.: Do respondents using smartphones produce lower quality data? Evidence from the first large-scale UK mixed-device survey – Understanding Society Wave 8, *International Journal of Social Research Methodology*, Vol. 28, No. 4, pp. 435–448 (online), DOI: 10.1080/13645579.2024.2397474 (2025).

[17] Moriarty, J.: *Qualitative Methods Overview*, SSCR Methods Reviews, National Institute for Health Research School for Social Care (2011).

[18] Nakayama, M. and Wan, Y.: The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews, *Information & Management*, Vol. 56, No. 2, pp. 271–279 (online), DOI: https://doi.org/10.1016/j.im.2018.09.004 (2019). Social Commerce and Social Media: Behaviors in the New Service Economy.

[19] Neigel, P., Selby, D. A., Arai, S., Tag, B., van Berkel, N., Vollmert, S., Vargo, A. and Kise, K.: Exploring the Alignment of Perceived and Measured Sleep Quality with Working Memory Using Consumer Wearables, *The Augmented Humans International Conference 2025*, AHs '25, New York, NY, USA, Association for Computing Machinery, (online), DOI: https://doi.org/10.1145/3745900.3746104 (2025).

[20] Nikolov, D., Oliveira, D. F., Flammini, A. and Menczer, F.: Measuring online social bubbles, *PeerJ Computer Science*, Vol. 1, p. e38 (online), DOI: 10.7717/peerj-cs.38 (2015).

[21] Nolasco, H. R., Vargo, A., Moreeuw, M., Hara, T. and Kise, K.: Augmenting Sleep Behavior with a Wearable: Can Self-Reflection Help?, *Proceedings of the Augmented Humans International Conference 2024*, AHs '24, New York, NY, USA, Association for Computing Machinery, p. 278–281 (online), DOI: 10.1145/3652920.3653049 (2024).

[22] Orth, U. R., Nickel, K., Böhm, R. and Röwe, K.: Agreement by design: The effect of visual harmony on responses to surveys, *Journal of Consumer Behaviour*, Vol. 19, No. 2, pp. 196–207 (2020).

[23] Ponnada, A., Wang, S. D., Li, J., Wang, W.-L., Dunton, G. F., Hedeker, D. and Intille, S. S.: Longitudinal User Engagement with Microinteraction Ecological Momentary Assessment (μEMA), *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 9, No. 3 (online), DOI:

10.1145/3749541 (2025).
[24] Russell, J., Weiss, A. and Mendelsohn, G.: Affect Grid: A Single-Item Scale of Pleasure and Arousal, *Journal of Personality and Social Psychology*, Vol. 57, pp. 493–502 (online), DOI: 10.1037/0022-3514.57.3.493 (1989).
[25] Trivedi, M. H.: Tools and strategies for ongoing assessment of depression: a measurement-based approach to remission, *The Journal of clinical psychiatry*, Vol. 70, No. suppl 6, p. 21046 (2009).
[26] van de Vijver, F. and Tanzer, N. K.: Bias and equivalence in cross-cultural assessment: an overview, *European Review of Applied Psychology*, Vol. 54, No. 2, p. 119–135 (online), DOI: 10.1016/j.erap.2003.12.004 (2004).
[27] van Herk, H., Poortinga, Y. H. and Verhallen, T. M. M.: Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries, *Journal of Cross-Cultural Psychology*, Vol. 35, No. 3, p. 346–360 (online), DOI: 10.1177/0022022104264126 (2004).
[28] Vargo, A. W. and Tag, B.: Using Peer-Production to Foster Bias Awareness among Online Content Consumers, *2020 CHI Workshop on Detection and Design for Cognitive Biases in People and Computing Systems*, (online),
[29] Weijters, B. and Baumgartner, H.: Misresponse to Reversed and Negated Items in Surveys: A Review, *Journal of Marketing Research*, Vol. 49, No. 5, pp. 737–747 (online), DOI: 10.1509/jmr.11.0368 (2012).