WORKING PAPER SERIES

BiTURF: Quantifying uncertainty to enhance strategic decision-making

Joshua Benjamin Schramm/Felix Josua Lang/Marcel Lichters

Working Paper No. 16/2026

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

FACULTY OF ECONOMICS
AND MANAGEMENT

Impressum (§ 5 TMG)

Herausgeber:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Wirtschaftswissenschaft
Der Dekan

Verantwortlich für diese Ausgabe:
J.B. Schramm, F. J Lang, M. Lichters
Otto-von-Guericke-Universität Magdeburg
Fakultät für Wirtschaftswissenschaft
Postfach 4120
39016 Magdeburg
Germany

http://www.fww.ovgu.de/femm

Bezug über den Herausgeber
ISSN 1615-4274

BiTURF: Quantifying uncertainty to enhance strategic decision-making

Joshua Benjamin Schramm[a]

joshua.schramm@ovgu.de

ORCID: 0000-0001-5602-4632

Felix Josua Lang[a]

felix.lang@ovgu.de

ORCID: 0000-0001-8572-4283

Marcel Lichters[a]*

marcel.lichters@ovgu.de

ORCID: 0000-0002-3710-2292

[a]Otto von Guericke University Magdeburg,

Faculty of Economics and Management,

39106 Magdeburg, Saxony-Anhalt, Germany

*Corresponding author

Acknowledgements

The authors would like to thank several anonymous editors and reviewers for their great comments on how to improve this manuscript. Furthermore, the authors highly appreciate Daniel Guhl's foundational work on the R/Stan model applied in the analyses as well as his helpful comments on a prior version of this manuscript.

Schramm, B. J., Lang, F. J., & Lichters, M. (2026). BiTURF: Quantifying uncertainty to enhance strategic decision-making. *Food Quality and Preference*, *145*, Article 106001. https://doi.org/10.1016/j.foodqual.2026.106001

BiTURF: Quantifying uncertainty to enhance strategic decision-making

Abstract

Total Unduplicated Reach and Frequency (TURF) analysis is widely used in both sensory product and market research for determining optimal product assortments under externally imposed constraints (e.g., a limited product development budget or a retailer's shelf space). However, researchers and practitioners face methodological limitations when applying classical TURF analysis, including the lack of quantification of uncertainty in results or issues with sparse data. To overcome these methodological limitations, this work introduces BiTURF—TURF analysis enriched with Bayesian input data—as a proof-of-concept and compares it with classical TURF analysis using two exemplary data sets. This work thereby demonstrates how to apply BiTURF to data from studies that use anchored Maximum Difference Scaling (i.e., Best-Worst Scaling) and Check-All-That-Apply data. In a nutshell, BiTURF quantifies the uncertainty of reach and frequency estimates and enables further analyses, including probabilistic head-to-head comparisons. This information provides researchers and practitioners with a richer foundation for their decisions. Furthermore, BiTURF overcomes the methodological corset of classical TURF, which can't be used with the *weighted by probability* approach on Check-All-That-Apply data. The article also provides a step-by-step *R* tutorial and concludes with a discussion of BiTURF's limitations and future research endeavors.

Keywords: Bayesian statistics; Check-all-that-apply; MaxDiff; Product assortment; TURF; Uncertainty

1 Introduction

Companies strive to find the best product assortment to attract customers and maximize revenue. However, they often face limitations in this endeavor. Imagine a beverage company that plans to launch a portfolio of iced teas. The company considered six flavors, but due to production costs and the limited shelf space retailers allowed, decided to develop and market only three. In other instances, the company might deliberately introduce only three variants to avoid choice overload for consumers (Hasted, 2023, p. 361). Likewise, for cost reasons, the company might be compelled to reduce its formerly larger product range to three variants in a product elimination decision. The standard method for identifying the optimal product combination under these constraints is the Total Unduplicated Reach and Frequency (TURF) analysis. TURF analysis, originally derived from media research (Miaoulis et al., 1990; Serra, 2013), nowadays belongs to the basic repertoire of many companies and leading software providers (e.g., Displayr, 2025; quantilope, 2024; Sawtooth Software, Inc., 2025a).

In the case of the imaginative beverage company, TURF analysis identifies the most promising combination of three out of the $\binom{6}{3} = 20$ possible flavor combinations based on consumers' preference data according to two metrics: *reach* and *frequency*. Reach is defined as the percentage of consumers who will buy at least one product from an assortment, while frequency (sometimes referred to as *volume*) is defined as how many products a consumer purchases on average from that assortment (Hasted, 2023, p. 363; Miaoulis et al., 1990).

TURF analysis has been applied in many different contexts, for example, to identify ideal combinations of flavored multi-vitamin/mineral gummies (Kuesten & Bi, 2021), to determine the optimal attributes of a hotel shampoo (Payne et al., 2023), or to search for the right combination of marketing actions for online apparel retailers (Schreiner & Baier, 2021). Besides its suitability for a broad range of business questions, one of TURF analysis' central benefits is its applicability to various sources of preference data resulting from popular market

research methods (Miaoulis et al., 1990). For example, Schramm and Lichters (2025a) applied TURF analysis to consumers' utilities derived from a *Maximum Difference Scaling study* (*MaxDiff*; also known as *Best-Worst Scaling*) on video games. Furthermore, Kuesten and Bi (2021) applied TURF analysis in conjunction with *Check-All-That-Apply* (CATA) questions, which are commonly used in consumer and sensory research (e.g., Ares & Jaeger, 2014; Bi & Kuesten, 2025).

However, currently available (hereafter *classical*) TURF analysis procedures have severe methodological limitations. On the one hand, although classical TURF analysis yields reach and frequency point estimates, it does not quantify the uncertainty of those parameters (i.e., estimation imprecision). Being unaware of or ignoring the uncertainty in the estimation can lead to the selection of suboptimal product assortments. In contrast, the proposed BiTURF procedure quantifies uncertainty as credible intervals for range and frequency estimates and provides additional information for decision-making, such as probabilistic head-to-head comparisons of competing product assortments.

On the other hand, classical TURF analysis prevents researchers from reaping the full potential of collected consumer data. In the case of MaxDiff, past research has applied TURF analysis, relying solely on point estimates of individual-level utility parameters, such as those provided by leading market research software (e.g., Schramm & Lichters, 2025a; Schreiner & Baier, 2021). However, because Bayesian modeling, commonly used with MaxDiff choice data in these software solutions, involves an iterative estimation process, one can instead use the information-rich posterior draws to portray uncertainty. Regarding CATA data, certain classical TURF approaches, such as the *weighted by probability* approach (see, e.g., Howell, 2016) can't currently be applied due to the binary nature of the CATA input data (e.g., purchase vs. non-purchase decisions).

To overcome the limitations of classical TURF analysis, this research proposes an advanced TURF approach, enhanced by parameter-uncertainty information, that uses a Bayesian input-generation procedure. Bayesian statistics rise in popularity due to widely discussed advantages (see, e.g., Etz & Vandekerckhove, 2018; Lagerkvist et al., 2012; Wagenmakers et al., 2018). These include (1) the possibility to estimate individuals' parameters taking knowledge from the whole sample into account (Veenman et al., 2024), (2) functionality with sparse data, a common problem in market and sensory product research relying on surveys or experiments (Karvanen et al., 2014; Lagerkvist et al., 2012), (3) the option to take prior knowledge into account via specification of priors, and (4) the opportunity to sample from the estimated posterior distributions to quantify uncertainty in the 'true estimate' (Wagenmakers et al., 2018). Specifically, in the present context, we demonstrate how to apply Bayesian-input TURF (BiTURF) to (anchored) MaxDiff and CATA data, given the relevance of these methods in market and sensory consumer research (e.g., Ares & Jaeger, 2014; Schuster et al., 2024). Herein, the demonstrated BiTURF analysis provides researchers and practitioners not only with the most promising product assortments based on reach and frequency but also goes beyond classical TURF results by quantifying the uncertainty in these parameters. This is achieved by leveraging posterior draws from a hierarchical Bayesian multinomial logit (HB-MNL) model. In addition, we demonstrate how to use an HB logit model to analyze CATA data, enabling the application of the *weighted by probability* TURF approach to these data as well. Regarding the *weighted by probability* approach, we also show how to obtain frequency estimates for product assortments, which are currently not provided by some proprietary software solutions.

The remainder of this article provides a brief introduction to the foundations of MaxDiff and CATA, as well as classical TURF analysis. Thereafter, it introduces BiTURF analysis. Subsequently, the proposed procedure—in comparison to classical TURF analysis—is demonstrated, drawing on data from both an anchored MaxDiff (Schramm & Lichters, 2025a)

and CATA (Kuesten & Bi, 2021). The present work additionally provides all *R* analysis scripts in an open-source repository, along with a step-by-step tutorial (Schramm et al., 2026). This tutorial also explains how to adapt BiTURF to other research scenarios and provides guidance on significantly reducing computational demands during estimation when computation time is a constraint in the research process.

2 Theory

2.1 Obtaining preference data for TURF analysis with MaxDiff or CATA

To enable a TURF analysis, researchers and practitioners must first collect preference data from consumers that allows for the classification of products (or more broadly items) into two classes: purchase options (or relevant items) and non-purchase options (i.e., irrelevant items). In a business research context, various question formats enable the collection and transformation of data into suitable formats. Two popular ways to do so are MaxDiff studies (see, e.g., Schramm & Lichters, 2025a) and CATA questions (see, e.g., Ares et al., 2014; Ares et al., 2011).

2.1.1 MaxDiff

The basis of a MaxDiff study to assess consumer preferences is a list of items (e.g., different product variants), which are presented to participants in a series of MaxDiff choice tasks (Louviere et al., 2013; Naughton et al., 2025). As illustrated in Step 1 of Fig. 1, a single MaxDiff choice task contains a subset of these items (usually three or more; Schuster et al., 2024), and participants must choose the alternative they would be most likely to buy and the alternative they would be least likely to buy (Louviere et al., 2013; Schramm & Lichters,

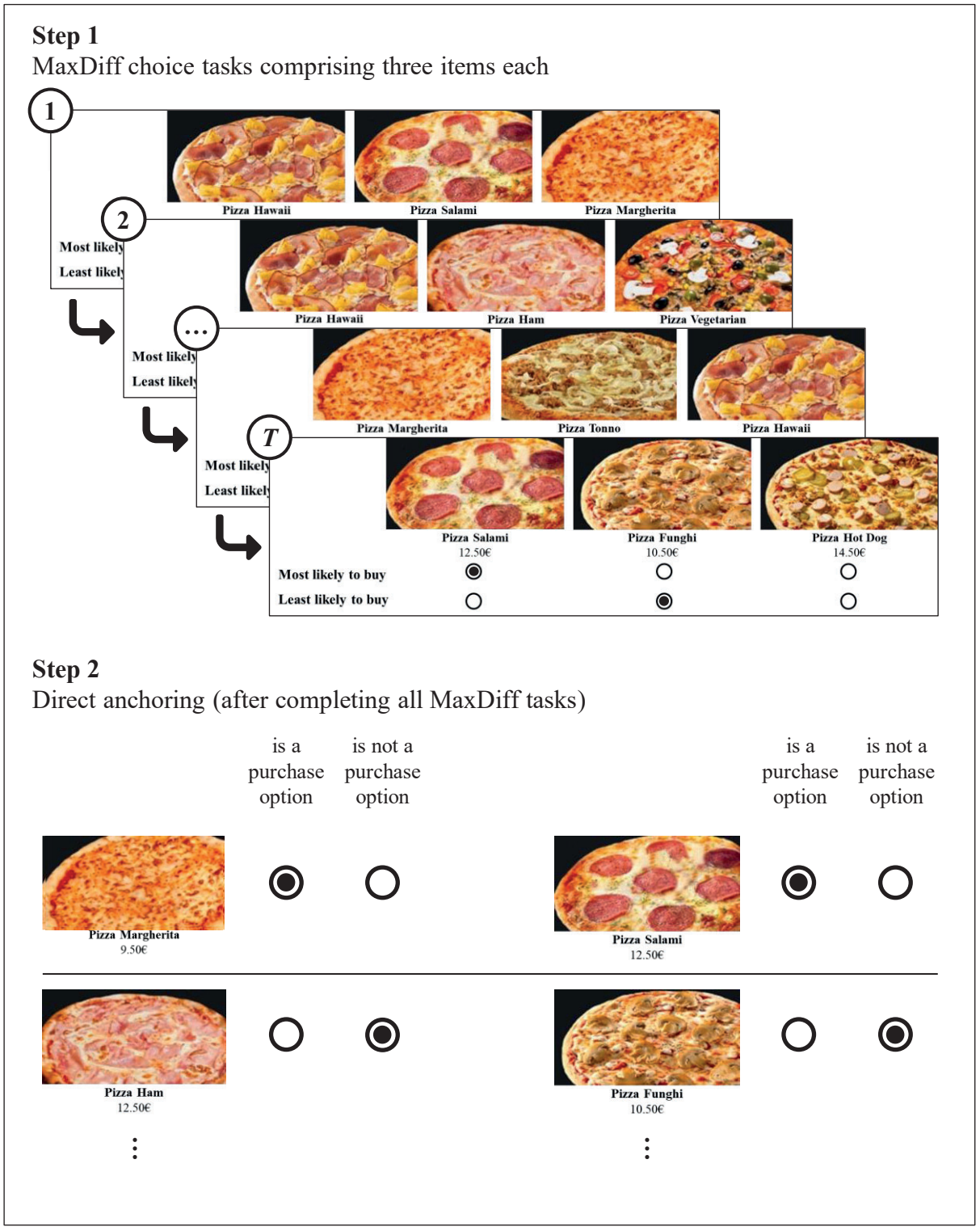

Fig. 1. Exemplary MaxDiff study including direct anchoring based on stimuli by Sablotny-Wackershauser et al. (2024).

2025a).[1] Responses to such MaxDiff choice tasks entail relative preferences in the pairwise comparisons of the choice tasks' items, that is, which item is preferred over another.

To allow for preference comparisons between participants (e.g., in terms of the absolute threshold for buying or not buying), researchers have implemented *anchored* MaxDiff variants, of which the two best-known variants are the direct and the indirect anchor approaches (Lagerkvist et al., 2012; Lattery, 2010; Orme, 2009). In direct anchored MaxDiff studies (see Step 2 of Fig. 1), participants indicate for all (or just some) of the items whether each is a purchase option or not after completing all MaxDiff tasks (Chrzan & Orme, 2019, pp. 86–87; Lattery, 2010). Therefore, anchoring provides a binary indication of whether products or attributes exceed the purchase threshold, that is, whether a participant would buy an individual product or not (see the Appendix for a more detailed description of a MaxDiff study, direct anchoring, and the associated coding in the design matrix before estimation). The direct anchor approach appears to be more common in practice; therefore, the present study focuses on direct anchoring as well (the accompanying *R* script, however, also works with data derived from an indirect anchored MaxDiff; Chrzan & Orme, 2019, p. 90, or other anchoring variants, see Schramm et al., 2025).

2.1.2 CATA questions

In a CATA question, as the name suggests, participants encounter items presented as a list and must check “all that apply” for purchase (Jaeger et al., 2015). Referring back to the introductory iced tea example, researchers could implement a CATA question after a sensory product acceptance test (Lichters et al., 2021) by displaying all six flavors to participants and asking them to indicate which flavors they would (and would not) purchase after tasting them.

[1] If the items are not products but, for example, product features, other wordings such as most important vs. least important are frequently used (for an overview, see, e.g., Schuster et al., 2024).

CATA questions, therefore, provide dichotomous data: a ‘1’ (or TRUE) indicates a purchase option, while a ‘0’ (or FALSE) indicates an option that isn't worth purchasing from a consumer's perspective. Participants usually consider CATA questions as easy (see, e.g., Jaeger et al., 2020) because, unlike MaxDiff, they view the stimuli only once and do not have to trade off among item alternatives.

2.2 Modeling preferences: From choices to utility parameters

After obtaining raw preference data, choices are often converted into a “usable metric”. In the case of a (anchored) MaxDiff study where preference data comprise multiple individual trade-offs between items (and in relation to the anchor), the choice data are initially coded in a design matrix (see Table A2 in the Appendix for an example of the best-worst coding approach that we used; Chrzan & Orme, 2019, pp. 20–22).[2] Next, researchers and practitioners typically implement an HB-MNL model to estimate participants' (part-worth) utilities for the incorporated items, which are interval-scaled preference parameters (see, e.g., Allenby et al., 2005). The anchoring allows utilities to be rescaled relative to the purchase threshold (e.g., utilities above vs. below the threshold; the anchor's utility is often set to 0), enabling inter-individual comparisons of willingness-to-buy.

HB-MNL models estimate utilities iteratively, yielding *posterior distributions* of these parameters relying on both sample- and individual-level information. These models are both standard in industry-leading software for MaxDiff and other preference elicitation techniques (see, e.g., Orme, 2000, Orme, 2021) as well as frequently applied in research, including the

[2] There are other coding approaches, such as MaxDiff or sequential coding, which would also be suitable for subsequent utility modeling and BiTURF analysis despite potential variations in resulting absolute parameter estimates as described by Chrzan and Orme (2019, pp. 25-26). The present paper relies on best-worst coding due to its widespread implementation in standard software such as Sawtooth Software, Inc. (2025a) or *R* packages as described by Aizaki and Fogarty (2023).

domain of food preferences (e.g., Jürkenbeck & Spiller, 2021; Klopčič et al., 2020; Naughton et al., 2025; Price et al., 2016). In the present context, HB-MNL models are generally suitable for providing input data to BiTURF due to their incorporation of estimates of individual utility parameters and distributions as a foundation for uncertainty quantification (see section 3.2). Furthermore, HB-MNL models account for the inter-personal preference heterogeneity (see, e.g., Allenby & Ginter, 1995; Lagerkvist et al., 2012; Veenman et al., 2024) and perform comparably well with sparse data (see, e.g., Allenby & Ginter, 1995; Allenby et al., 2005; Lenk et al., 1996; Orme & Chrzan, 2021, p. 145). HB-MNL is one of many approaches to account for heterogeneity (see also Orme & Chrzan, 2021, pp. 156–157). There are other viable modeling procedures, such as mixed logit models or multinomial probit models. However, in the case of the former, individual utility parameters would have to be extracted post-hoc or approximated through simulations (Elshiewy et al., 2016; Elshiewy et al., 2017), and the procedure is less suitable for sparse data (Allenby et al., 2005; Orme & Chrzan, 2021, p. 145), while the latter has higher computational demands (see, e.g., Loaiza-Maya & Nibbering, 2022). Also, both models are often not estimable with sparse data.

To ensure data quality of estimated utilities, MaxDiff studies should be designed and tested upfront accordingly (for detailed guidelines, see Chrzan & Orme, 2019). Generally, each participant should face each item at a minimum of three to four times throughout the MaxDiff study to obtain stable individual preference estimates. This can be achieved through sufficient items per MaxDiff task and the number of MaxDiff tasks per participant. Regarding the choice set size, a higher number of items (as a rule of thumb, four to five) provides more pairwise comparisons per choice task (Schuster et al., 2024). The number of choice tasks should then ensure the suggested exposure to each item. Observing insufficient choices per individual results in shrinkage, meaning that individual utilities gravitate stronger towards the sample's mean (see, e.g., Orme & Chrzan, 2021, pp. 156–157; Train, 2009, pp. 266–267).

In contrast to a MaxDiff study, participants in a CATA study respond to only one CATA question, indicating purchase relevance for each item. Although the data format's simple, binary nature allows straightforward interpretation and can be used in subsequent analyses, such as classical TURF, without modeling or transformation (see section 2.3), the data are naturally sparse. Therefore, we propose to apply Bayesian preference modeling to CATA data, which utilizes information shared across participants to obtain enriched results, compensating for choice error (Allenby et al., 2005), and, central to this present article, to allow for BiTURF analysis. That is, we apply an HB logit model to fit the binary nature of CATA (see section 3.1); however, other models could also be implemented (e.g., multivariate binomial probit model; Edwards & Allenby, 2003).

2.3 TURF analysis

2.3.1 Classical TURF analysis

Independent of the previously described data-collection methods, TURF analysis can help determine product assortments that maximize revenue based on reach and frequency. Let's revert to the initial example of the beverage company to illustrate the concept of TURF analysis (see also Serra, 2013). Assume the company asked eight consumers to indicate each flavor they would purchase (i.e., CATA). Given the results in Table 1, the company would decide on the three-flavor combination of *Peach*, *Citrus*, and *Wild Berry*. This combination is the only assortment of which every consumer would buy at least one product (reach = 100%). Furthermore, this assortment would yield an expected total of 12 products sold to the eight consumers (frequency = 1.50). Note that for small problems like this, one can easily arrive at an optimal decision. However, this is no longer an option in typically applied research settings when expanding the number of products or interviewing hundreds of consumers.

Table 1
Exemplary input data for TURF analysis on iced tea.

	Cherry	**Peach**	Melon	**Citrus**	**Wild Berry**	Green Tea
Consumer 1	1	**0**	0	**0**	**1**	0
Consumer 2	1	**0**	0	**1**	**0**	0
Consumer 3	0	**1**	0	**0**	**1**	0
Consumer 4	0	**0**	1	**0**	**1**	1
Consumer 5	0	**1**	0	**0**	**0**	0
Consumer 6	1	**1**	1	**1**	**1**	1
Consumer 7	0	**0**	0	**1**	**0**	0
Consumer 8	0	**1**	0	**0**	**1**	0

Note: This example applies the *threshold* approach to CATA data. The numbers in the cells indicate that the participant would buy (= 1) or not buy (= 0) the respective product.

2.3.2 TURF ladder

A TURF ladder (also known as an incremental plot) is a popular presentation form for TURF results (e.g., Hasted, 2023, p. 363). While the total assortment size is fixed as a restriction for TURF analysis, the TURF ladder starts with a single-item solution (i.e., the iced tea with the highest reach) and gradually adds the item with the next-highest incremental reach to the existing assortment. Referring to Table 1, the results show that *Wild Berry* should be introduced first because it has the highest reach of all the flavors (62.50%). Next, *Citrus* would be added, attracting two new consumers and thereby increasing the total reach to 87.50%. Note that *Cherry* or *Peach* at this stage would only attract one new consumer despite having an equal (i.e., three) or higher (i.e., four) individual reach in relation to *Citrus* because *Wild Berry* already attracts the majority of the consumers that *Cherry* or *Peach* would attract (two and three, respectively). However, *Peach* is added to the assortment as the final third product to achieve a 100% reach (i.e., also attracting consumer 5). Fig. 2 presents the corresponding TURF ladder. The last three items only increase the frequency values for the assortment and would be added in the following order: *Cherry*, *Melon*, and *Green Tea*.[3]

[3] Kuesten and Bi (2021) used a slightly different approach in which choices from previous assortment sizes are not necessarily part of the larger combinations. We stick to the proposed approach above because it is more common in market research software such as quantilope (2024) or Displayr (2025).

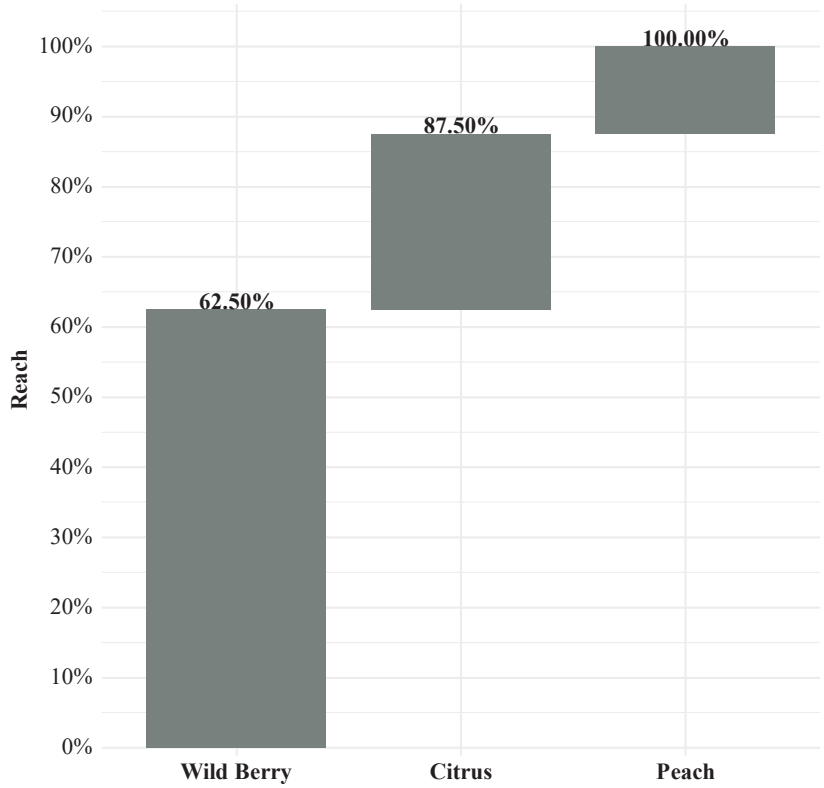

Fig. 2. TURF ladder for the iced tea example, indicating the (incremental) reach.

2.3.3 Existing TURF approaches

The literature proposes various approaches to TURF analysis (for an overview, see Chrzan & Orme, 2019, pp. 111–113). This research focuses on the *threshold* and the *weighted by probability* approaches, as they are most often implemented in off-the-shelf market research software. The *threshold* approach is illustrated in the example above and analyzes preference data by considering all items with a utility above a defined threshold (i.e., the purchase threshold) in a binary format (i.e., all product alternatives marked as purchase options). Because of the binary data nature, the *threshold* approach is the only TURF approach currently available for CATA data (e.g., as done in Kuesten & Bi, 2021). To conduct a TURF analysis using the *threshold* approach based on anchored MaxDiff data, metric utilities are recoded as binary values. Utilities exceeding the threshold's (i.e., no-buy alternative's) utility are assigned a ‘1’, and ‘0’ otherwise (see, Schramm & Lichters, 2025a, for a market research example).

However, recoding interval-scaled MaxDiff utilities to discrete values discards information and is problematic (Howell, 2016). More specifically, when the utility coefficients' metric scale level derived from an HB-MNL model is converted to a binary format, the coefficients' magnitudes lose importance. For example, assume the anchor (i.e., no-buy utility) is set to '0' and consumer 1 has a utility of 2.35 for *Wild Berry* and a utility of 0.10 for *Cherry*. In the *threshold* approach, the recoded data (1 in both cases) imply that each flavor is equally likely to be purchased. However, consumer 1 actually prefers *Wild Berry* to *Cherry*-flavored iced tea. To benefit from the metric scale level of utilities derived from MaxDiff studies and to circumvent the problems of data reductionism, Sawtooth Software implemented the *weighted by probability* approach (Chrzan & Orme, 2019, p. 112; Howell, 2016). Therein, the MaxDiff raw utility scores are used to estimate the probability that a participant chooses an item over the no-buy alternative, thereby delivering more precise parameter estimates (Chrzan & Orme, 2019, p. 112; Howell, 2016). For example, the TURF analysis reported in Schreiner and Baier (2021) applied the *weighted by probability* approach. However, in contrast to an estimate of reach, Sawtooth Software, Inc. (2025b) currently does not provide frequency estimates. To obtain these parameters, we propose calculating frequency estimates as the expected number of purchased items. Specifically, frequency is the sum of respective probabilities that a participant chooses an item over the no-buy alternative across all considered items (averaged across participants).

Although the *weighted by probability* approach goes beyond the *threshold* approach by using metric purchase probabilities, it is not possible to quantify the uncertainty in reach and frequency estimates within classical TURF. This is because, as seen in previous papers, researchers who applied a TURF analysis on MaxDiff results use a single aggregated utility estimate per participant as input parameters (i.e., beta point estimate calculated by aggregating all posterior draws from the HB-MNL model; see, e.g., Payne et al., 2023; Schramm &

Lichters, 2025a; Schreiner & Baier, 2021), which prevents assessing the uncertainty inherently provided in Bayesian posterior sampling. This is unsatisfactory, as it alternates between Bayesian estimates of the MaxDiff data and a frequentist TURF analysis based on singular point estimates.

3 Material and methods

In contrast to classical TURF analysis, applying Bayesian statistics to process the input data entering TURF analysis enables assessments of the uncertainty in reach and frequency estimates. Specifically, our proposed approach provides uncertainty estimates for TURF parameters by sampling from the Bayesian posterior distributions of utility estimates derived from both MaxDiff and CATA data.

3.1 Data sets

To illustrate BiTURF, the remainder of the present paper draws on two published data sets.[4] First, Schramm and Lichters (2025a) provide an incentive-aligned direct-anchored MaxDiff data set ($n = 112$; screened for attention, speeding, straightlining, and individual model fit based on root likelihood; Orme, 2019). This research is about 16 PlayStation 5 games.[5,6] Participants in their study selected their most and least preferred items from sets of

[4] We thank both author groups for making the data available.

[5] We also want to refer the reader to the ErrVarNorm criterion as another approach to assess data quality in MaxDiff studies, as proposed by Jaeger et al. (2026a), Jaeger et al. (2026b), and Llobell et al. (2025). To control for extreme response behavior, Gensler et al. (2012), Sablotny-Wackershauser et al. (2024), and Schramm et al. (2025) serve as good starting points for aspiring users.

[6] We opted to use this data set for multiple reasons. First, the data serves as an openly available, generic input for the proof-of-concept illustration explained in this paper and the tutorial. Second, there are few available (incentive-aligned) direct-anchored MaxDiff data sets that allow for demonstrating BiTURF with reasonable parameter estimates.

four, in a total of 16 consecutive MaxDiff tasks. The direct anchor questions followed after completion of the MaxDiff tasks (see section 3.1 of their article). Given this design, the sample size allows for estimating utility parameters with satisfying precision (Lipovetsky et al., 2015).

Second, as an illustrative CATA data set, the data published by Kuesten and Bi (2021; see section 3.4 of their article; $n = 200$) is employed.[7] Here, consumers selected functional ingredients of a multi-vitamin/mineral gummy product. The study included a total of nine ingredients.

3.2 Bayesian input data and BiTURF analysis

To obtain Bayesian input data for subsequent BiTURF analysis, we apply HB modeling to both the MaxDiff and CATA raw choice data to estimate preference parameters. The analysis of MaxDiff data employs best-worst coding (Chrzan & Orme, 2019, pp. 20–22) and applies the HB-MNL model provided by Schramm et al. (2025), relying on a single multivariate normal prior (Allenby & Ginter, 1995). The analysis of the CATA data uses an HB logit model because the data are binary. Both cases involved five independent MCMC (Markov chain Monte Carlo) chains, each with 4,000 iterations, including 1,000 warm-up iterations, and a thinning parameter of 5. Thus, the models retain 3,000 ($5 \cdot \frac{4{,}000-1{,}000}{5}$) posterior draws of utilities for each individual. Model estimation relied on *R*/*Stan* (Carpenter et al., 2017), implementing weakly informative priors. Both models show satisfactory convergence and efficiency statistics ($\hat{R}$ values all below 1.01, effective sample sizes above 1,000, and no divergent transitions during estimation; Carpenter et al., 2017; Stan Development Team, 2025). Furthermore, internal fit statistics and predictive accuracy scores suggest good model fit in

[7] Kuesten and Bi (2021) do not report any screening measures applied on their sample (except for quota sampling during data collection). We abstained from any screening to keep the data in line with the original study and because of missing behavioral data that would allow for applying exclusion criteria like speeding.

both cases (see Tables A4 and A5 in the Appendix).[8] The accompanying *R* tutorial provides more detailed information on data processing and the respective modeling approaches.

Subsequently, the actual BiTURF analysis uses the respective set of posterior draws to identify dominant product assortments and quantify their associated uncertainty. Specifically, a TURF analysis is conducted on the utilities from each of the 3,000 iterations to obtain point (i.e., mean) and uncertainty (i.e., 95%-credible intervals) estimates for the reach and frequency parameters. The analysis, thereby, draws on the *R* package *validateHOT* (Schramm & Lichters, 2025b) for both classical and BiTURF, extended by custom functions to implement the *weighted by probability* approach. We present the results of both the classical TURF and BiTURF analyses below; the BiTURF results serve primarily as a proof-of-concept demonstration. The tutorial provided in the corresponding OSF repository (Schramm et al., 2026) illustrates how to adjust BiTURF to fit additional application contexts (e.g., different sizes of product assortments).

4 Results

4.1 Anchored MaxDiff

For comparison, we first run a classical TURF analysis using the aggregated beta-point utility estimates.[9] Schramm and Lichters (2025a) provided a combination of three products; therefore, we also focus on this assortment size. Table 2 displays the results of the *threshold* and *weighted by probability* approaches for both the classical and BiTURF analyses.

[8] The predictive accuracy measure reported in the Appendix assesses how accurately the posterior draws predict actual choices. Note that conventional posterior predictive checks (Gelman et al., 1996) assess predictive accuracy by comparing observed data with simulated outcomes generated from the posterior distribution (typically based on comparisons of graphical diagnostics or summary statistics; Carpenter et al., 2017; Gelman & Hill, 2007, p. 513; Johnson et al., 2022).

[9] Note that participants' estimated utilities deviate from those by Schramm and Lichters (2025a) because they used Sawtooth Software (2025a) instead of *R*/*Stan* for model estimation.

Table 2
Classical vs. BiTURF (*threshold* and *weighted by probability* approach, MaxDiff).

	#	Combination	Reach[1]	Frequency[1,2]	Rank[1]	% first[3]
Classic Threshold	1	G5, G7, G15	89.29%	1.44	/	/
	2	G5, G8, G15	89.29%	1.44	/	/
	3	G5, G8, G13	89.29%	1.35	/	/
	4	G4, G5, G15	88.39%	1.58	/	/
	5	G4, G10, G15	88.39%	1.53	/	/
Classic WBP	1	G5, G7, G15	87.38%	1.46	/	/
	2	G7, G10, G15	86.76%	1.41	/	/
	3	G7, G11, G15	86.63%	1.43	/	/
	4	G4, G10, G15	86.34%	1.49	/	/
	5	G5, G8, G13	86.00%	1.33	/	/
BiTURF Threshold	1	G5, G7, G15	88.03% [83.93%, 91.96%]	1.45 [1.37, 1.54]	5.45 [1.00, 23.00]	28.17%
	2	G7, G10, G15	87.80% [83.93%, 91.96%]	1.41 [1.31, 1.49]	6.43 [1.00, 28.03]	21.83%
	3	G7, G11, G15	87.45% [83.04%, 91.07%]	1.42 [1.33, 1.50]	8.00 [1.00, 34.00]	15.00%
	4	G4, G10, G15	87.39% [83.04%, 91.07%]	1.49 [1.39, 1.57]	7.80 [1.00, 29.00]	15.77%
	5	G5, G8, G13	87.11% [83.04%, 91.07%]	1.32 [1.24, 1.41]	9.98 [1.00, 40.03]	20.03%
BiTURF WBP	1	G5, G7, G15	86.38% [84.05%, 88.58%]	1.46 [1.41, 1.52]	2.39 [1.00, 9.00]	52.87%
	2	G7, G10, G15	85.60% [83.18%, 88.00%]	1.42 [1.36, 1.48]	5.52 [1.00, 17.00]	9.40%
	3	G4, G10, G15	85.50% [83.17%, 87.76%]	1.48 [1.42, 1.54]	5.88 [1.00, 18.00]	11.27%
	4	G7, G11, G15	85.41% [82.91%, 87.80%]	1.43 [1.38, 1.49]	6.75 [1.00, 20.00]	5.10%
	5	G4, G6, G15	85.00% [82.47%, 87.34%]	1.50 [1.44, 1.56]	9.12 [1.00, 24.00]	4.77%

Note: WBP = *weighted by probability* | G = Generic product | [1]95% credible interval in brackets | [2]For WBP: Sum of participants' purchase probabilities of the assortments' items | [3]Percentage of posterior draws in which a combination was ranked first according to total reach (the better (i.e., lower) rank was assigned to the combinations in case of tied ranks).

In classical TURF analysis, the best 3-product assortment in terms of both reach and frequency comprises G5, G7, and G15 across both approaches. Two out of four of the second through fifth product assortments are identical across the *threshold* and *weighted by probability* approaches (i.e., [G4, G10, G15] and [G5, G8, G13]), but their respective ranks differ. Furthermore, the *weighted by probability* approach yields smaller reach estimates than the *threshold* approach.

BiTURF enriches the results by providing the 95%-credible intervals for reach and frequency estimates (i.e., the Bayesian probability intervals that contain the true values). Furthermore, one can extract the average rank of product assortments (including their 95%-credible interval; the lower the better) as well as the number of times each product assortment scored first across all posterior draws (i.e., *% first*; the higher the better). The two BiTURF approaches' results patterns are similar to those of classical TURF. Specifically, the first- and second-ranked product assortments are the same for the *threshold* and the *weighted by probability* BiTURF approaches, while the third- to fifth-ranked assortments differ (in terms of ranking and assortment composition). Furthermore, while the *weighted by probability* approach yields lower reach estimates than the *threshold* approach, it provides more precise estimates as reflected in smaller credible interval width. Additionally, the *weighted by probability* BiTURF showed more pronounced differences in how often the respective trinary assortments are ranked first across the posterior draws (see, e.g., the *% first* column). Note, however, that these are empirical patterns observed in this example. Although this is a promising result in favor of the *weighted by probability* approach, further confirmation and generalizability are necessary for firm conclusions.

Comparing the results of the classical and BiTURF analyses, both procedures yield the same best product assortment, but differ slightly in the product compositions and ranks of the second to fifth assortments within each approach. Furthermore, classical TURF tends to yield

descriptively higher reach scores for the top product assortments than BiTURF analysis in both the *threshold* and *weighted by probability* approaches, whereas frequencies differ more inconsistently across the procedures.

Let's think for a moment about a market researcher who uses the classical TURF with the *weighted by probability* approach. Based on the classical TURF's findings, the researcher might be relatively indifferent between recommending one of the top 3-product assortments, because the difference in reach between adjacent combinations is marginal, which means less than 1%. If the researcher thinks that the products G4 and G10 are easier to implement than G5, this might lead to a recommendation of even the fourth-best 3-product assortment. The point, however, is that BiTURF delivers an additional argument against doing so in this example, although the combination [G4, G10, G15] is even better ranked than in the classical analysis (third instead of fourth). More precisely, opting for the best 3-product assortment [G5, G7, G15] entails less uncertainty, as this combination is estimated to be much more likely to be the best recommendation than the other assortments (see *% first*). In other applications, the estimates for *% first* might rather support the researcher's decision not to opt for the best 3-product assortment (i.e., if estimates are closer to each other).

Implementing BiTURF analysis offers further benefits over its classical counterpart because it allows for assessing the posterior-based probability that one product assortment outperforms another (Bayesian *p*-values) in direct comparisons (see, e.g., Rooderkerk & Lehmann, 2021). More specifically, one can report the probability that one assortment is preferred over the other by examining how many of the 3,000 posterior draws one has a higher reach (or frequency) than the competing assortment. For example, comparing assortments #1 and #2 from the Bayesian *threshold* approach head-to-head, #1 [G5, G7, G15] dominates in reach in 44.90% of the draws over #2 [G7, G10, G15], which wins 35.17% of the draws. In the remaining draws, the two alternatives tie in terms of reach. When applying the *weighted by*

probability approach instead for this comparison, the results clearly favor the first combination with a higher reach in 79.80% (vs. 20.20%) of the posterior draws. These comparisons can provide further confidence in deciding between two assortments for product development or sales, or in favoring the one that is easier to implement if they do not perform significantly differently (i.e., if they approach the same win-loss percentages).

Next, Fig. 3 illustrates the respective TURF ladders. For all approaches, the order of items included in the product assortment is identical, except for the classical *threshold* approach (Panel A). More importantly, while the classical TURF ladder reports only the reach and its incremental growth when adding a further item, the BiTURF ladder again reports the corresponding credible interval and, in addition, the percentage of draws supporting this order. For example, according to the *weighted by probability* approach in Panel D, G5 should be introduced first, reaching on average 56.57% of the participants (95% CI = [54.53%, 58.80%]). In 54.37% of the 3,000 posterior draws, G5 has the highest reach and should be added first instead of any other product. For an assortment of size 2, G15 should be added next. This is supported by 24.27% of the 3,000 posterior draws. Alternatively, one could also calculate the number of draws supporting adding G15 second given that G5 was introduced first ($\frac{0.2427}{0.5437} \times 100 = 44.64\%$). The accompanying tutorial also includes the code to create the BiTURF ladders.

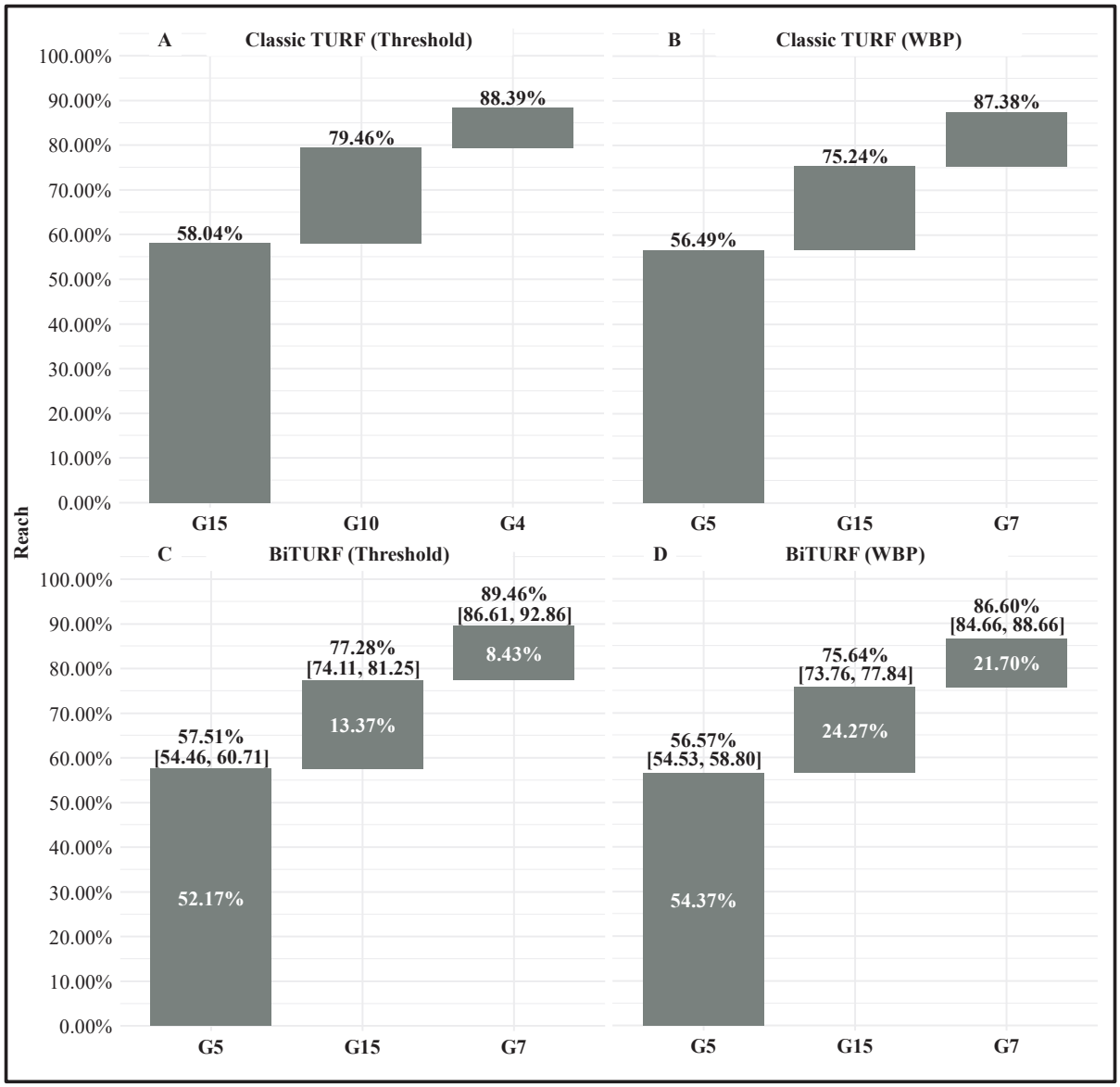

Fig. 3. TURF ladders (anchored MaxDiff) | G = Generic product | numbers above the bars indicate a product's (incremental) reach, including credible intervals for BiTURF, and numbers in the bars indicate the percentage of 3,000 posterior draws in which the product has the highest reach at the current addition step.

4.2 CATA

We first replicate the classical TURF results by Kuesten and Bi (2021; Table 3, *classic threshold*). The best combination contains supplements S1, S3, and S9. As discussed, the *weighted by probability* approach is not available in classical TURF because CATA's data are binary (e.g., purchase vs. no purchase).

Table 3
Classical vs. BiTURF – both for the *threshold* and the *weighted by probability* approach for CATA data.

	#	Combination	Reach[1]	Frequency[1,2]	Rank[1]	% first[3]
Classic Threshold	1	S1, S3, S9	83.00%	1.14	/	/
	2	S3, S7, S9	81.50%	1.10	/	/
	3	S3, S4, S9	80.50%	1.07	/	/
	4	S3, S6, S9	79.00%	1.03	/	/
	5	S1, S3, S4	78.50%	1.43	/	/
BiTURF Threshold	1	S1, S3, S9	79.77% [73.00%, 86.00%]	1.13 [1.03, 1.24]	1.91 [1.00, 6.00]	57.43%
	2	S3, S7, S9	78.38% [71.50%, 84.50%]	1.09 [0.99, 1.20]	3.09 [1.00, 10.00]	27.33%
	3	S3, S4, S9	77.03% [70.00%, 83.50%]	1.05 [0.94, 1.16]	4.67 [1.00, 14.00]	13.23%
	4	S1, S3, S4	75.82% [70.00%, 81.00%]	1.42 [1.30, 1.53]	5.97 [2.00, 14.00]	2.23%
	5	S3, S4, S7	75.53% [69.50%, 81.00%]	1.38 [1.26, 1.50]	6.53 [1.00, 15.00]	3.10%
BiTURF WBP	1	S1, S3, S9	77.55% [73.13%, 81.84%]	1.14 [1.07, 1.21]	1.96 [1.00, 7.00]	57.67%
	2	S3, S7, S9	76.40% [71.74%, 80.71%]	1.10 [1.03, 1.17]	3.55 [1.00, 10.00]	21.10%
	3	S1, S3, S4	75.43% [71.65%, 78.97%]	1.43 [1.35, 1.51]	5.00 [1.00, 10.00]	4.13%
	4	S1, S3, S7	75.26% [71.53%, 78.78%]	1.46 [1.38, 1.54]	5.33 [2.00, 10.00]	1.77%
	5	S3, S4, S9	75.24% [70.47%, 79.73%]	1.07 [1.00, 1.15]	5.63 [1.00, 13.00]	6.00%

Note: WBP = *weighted by probability* | S = supplement | [1]95% credible interval in brackets | [2]For WBP: Sum of participants' probabilities of the combinations' items | [3]Percentage of posterior draws in which a combination was ranked first according to total reach (the better rank was assigned to the combinations in case of tied ranks).

In contrast, BiTURF analysis supports both the *threshold* and the *weighted by probability* approaches. Again, we obtain credible intervals for both reach and frequency parameters, and we can also determine the average rank and the share of first placements for each supplement combination. The first- and second-ranked ingredient combinations are the same in both BiTURF approaches; however, the third- to fifth-ranked combinations differ. The *weighted by probability* approach yields lower reach estimates than the *threshold* approach, but these estimates are more precise (i.e., narrower credible intervals) in the present case, as in the MaxDiff results.

When comparing the classical and BiTURF analyses regarding the *threshold* approach, the three highest-ranked combinations are identical. However, both the estimated reach and frequency scores are lower in BiTURF, yielding more conservative estimates.

As with MaxDiff, Bayesian *p*-values for head-to-head comparisons of combinations of functional supplements can be obtained in BiTURF analysis applied to CATA data. Comparing the two best-ranked 3-item combinations in the Bayesian *threshold* approach, the first wins in 64.60% of draws, whereas the other would be ranked first in 29.47% of draws.

Finally, Fig. 4 shows the TURF ladder results for the classical *threshold* (Panel A), the BiTURF *threshold* (Panel B), and the BiTURF *weighted by probability* approach (Panel C). Again, the BiTURF ladders provide credible intervals for the (cumulative) reach. All approaches result in the same order of supplements to be added, yielding similar reach estimates in the *threshold* approaches, while the Bayesian *weighted by probability* approach provides the lowest point estimates, matching the results displayed in Table 3. However, the posterior probabilities supporting the item entry order are higher under the *weighted by probability* approach than under the *threshold* approach.

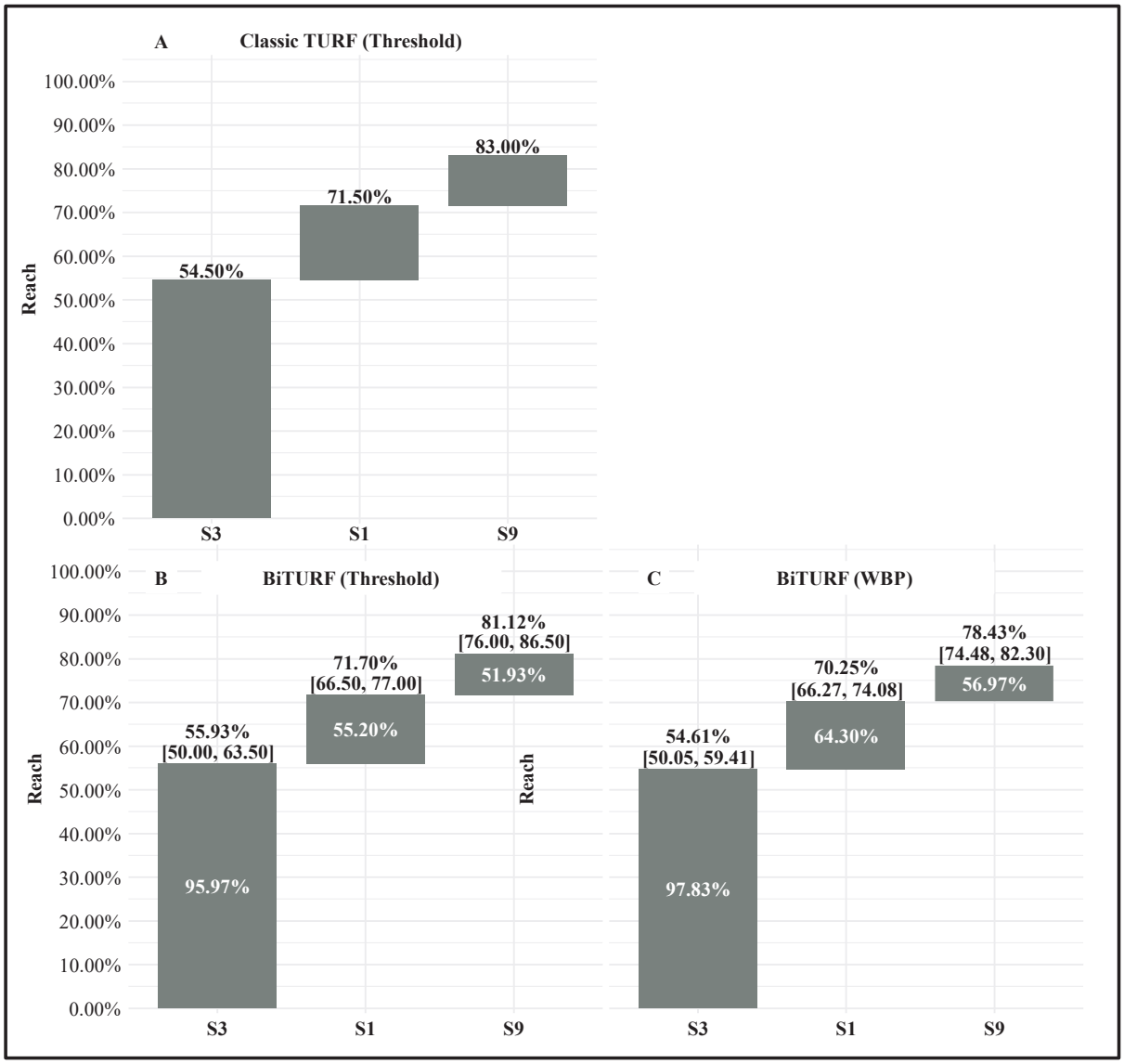

Fig. 4. TURF ladders for the CATA data | S = Supplement | numbers above the bars indicate a supplement's (incremental) reach, including credible intervals for BiTURF, and numbers in the bars indicate the percentage of 3,000 posterior draws in which the supplement has the highest reach at the current addition step.

5 Discussion

5.1 Summary and implications

Finding the most promising product assortment is of interest to both manufacturers and retailers, as they seek the product combination that maximizes their business success (Rooderkerk et al., 2013). One frequently used method for defining such assortments is TURF analysis. It has been applied to various questions and has the advantage of being applicable to various data sources, for example, preference data obtained from MaxDiff (Schreiner & Baier,

2021) or CATA questions (Kuesten & Bi, 2021). However, it was previously not possible to quantify the uncertainty in the results of classical TURF analysis—a gap that the present work closes. More specifically, the proposed BiTURF procedure uses Bayesian posterior sampling to portray uncertainty in TURF results. In essence, the use of BiTURF enriches decision-making by providing uncertainty estimates and several additional parameters beyond those of classical TURF. To facilitate advancements in TURF analysis, this article demonstrates how to apply BiTURF in a proof-of-concept case study using two exemplary data sets that rely on anchored MaxDiff and CATA formats (Kuesten & Bi, 2021; Schramm & Lichters, 2025a). It additionally provides a step-by-step *R* tutorial, allowing future researchers and practitioners using an open-source solution to fully benefit from BiTURF.

The presented case studies applied the *threshold* and *weighted by probability* approaches. Fundamentally, BiTURF analysis provides uncertainty estimates for the reach and frequency parameters, the focal results of TURF. Furthermore, while the classical TURF analysis only reports point estimates for reach and frequency, applying Bayesian statistics to the input data used in BiTURF analysis can also provide average ranks (including uncertainty intervals) and the share of first-ranked iterations of product assortments across posterior draws. Finally, since the results of TURF analysis (especially the top recommended product assortments) are often quite close to each other, we provide a way to examine the probability that one product assortment is superior to another (i.e., Bayesian *p*-values based on posterior probabilities in a head-to-head comparison). Taking into account the uncertainty in reach and frequency estimates, combined with these additional metrics, provides decision makers with richer information when using BiTURF than when using classical TURF analysis. This information helps quantify the confidence and risk involved in selecting a product assortment, for example, during a market launch, during product elimination, or when allocating shelf space. In addition, considering the uncertainty of parameter estimates and Bayesian *p*-values

in head-to-head comparisons can also help to justify selecting a product assortment that is not the best in terms of mean reach and frequency (e.g., if it is cheaper and easier to implement). In the demonstration's results, for example, the uncertainty in parameter estimates is comparatively large relative to the differences in these parameters across product assortments. That is, given the same retail prices, expected revenue differences across product assortments are, by definition, insignificant and could be easily offset by lower production costs.

As a further advancement, we provide a solution to apply the *weighted by probability* approach, and not just the *threshold* approach, in BiTURF analysis of CATA data. The former has the added benefit of accounting for the magnitude of utilities, a factor neglected by both the *threshold* and *first-choice* approaches.[10] This might potentially lead to more precise TURF results. Due to the binary nature of CATA data, applying the *weighted by probability* approach was not feasible in classical TURF analysis. The presented BiTURF solutions from the *weighted by probability* approach are more precise and consistent across posterior draws than those from the *threshold* approach for the exemplary MaxDiff and CATA data. Although not a general proof, this is a promising result and indicates that the *weighted by probability* approach might yield more trustworthy results for TURF parameters (Howell, 2016). Additionally, because leading software solutions do not currently always provide frequency estimates for the *weighted by probability* approach, we present a straightforward estimation procedure to obtain these parameters and thereby reap the full benefit of this TURF approach.

5.2 Limitations and future research

One general limitation of applying BiTURF analysis is that it requires more computation time than classical TURF. To speed up BiTURF analysis, we implemented

[10] The provided tutorial additionally allows running the *first-choice* rule in BiTURF; see Chrzan and Orme (2019, p. 111).

functions that enable parallel processing across multiple CPU cores. To further reduce computation time, we invite future researchers to implement other TURF approaches, such as eTURF (Ennis et al., 2012) or binary linear programming (Serra, 2013). In addition, we provide another alternative to reduce computation time in the accompanying tutorial by dividing the analysis into two steps. More specifically, we suggest finding a range of combinations using fewer posterior draws (i.e., increasing the thinning parameter, which still can be done after HB estimation) and then focusing on this subset again with more posterior draws to obtain more accurate results for the top combinations. This ignores combinations that are not promising from the outset and therefore unlikely to be implemented. Comparing the proposed two-step process (91.66 s) with the extensive BiTURF estimation (360.60 s) using the analyzed MaxDiff data, the former can potentially quarter the computation time (at least in this example).[11]

Another general limitation warrants emphasis. Although we compared the results of the Bayesian and classical TURF analyses, including both the *threshold* and *weighted by probability* approaches where possible, this comparison is not a *benchmark*. In other words, we do not know which TURF analysis, using which approach, most closely approximates the true parameter values. In this regard, it is also important to emphasize that the presented results reflect empirical patterns derived from two exemplary data sets. That is, the results (e.g., more precise parameter estimates in BiTURF's *weighted by probability* vs. *threshold* approach) should not be overgeneralized but rather be viewed as emerging from a methodological proof-of-concept case study in need of further corroboration. A simulation study might shed light on these topics in the future. Such a simulation study could also provide information on the influence of the MaxDiff design (e.g., the number of items in the MaxDiff, the number of items

[11] We ran the comparison ten times and obtained computation times using the *R* package *microbenchmark* (Mersmann et al., 2024). Runtime varies depending on the hardware and software; in the present case, a Lenovo ThinkCentre M70t G4 and *R* version 4.3.3.

per MaxDiff task, the anchoring method, and the number of MaxDiff tasks per participant) on the levels of uncertainty in parameter estimates.

Furthermore, the present work compared BiTURF results based solely on the posterior distribution from an HB model with those of classical TURF. However, BiTURF is not the only conceivable possibility to generate input data for uncertainty quantification based on parameter distributions. For example, future researchers could direct their inquiries to other techniques, such as comparing BiTURF with classical bootstrapping (Efron, 1986; Efron & Tibshirani, 1993), Bayesian bootstrapping (Rossi et al., 2024), or mixed logit.

In addition, while the *weighted by probability* in comparison to the *threshold* approach takes into account the magnitude of the coefficients, it falls prey to the issue of scale confound in logit models, defined as the magnitude of the preference utilities proportionate to the error term's size (see, e.g., Hauser et al., 2019). However, there is no clear indication of how scale influences parameter estimation in BiTURF. One can speculate that as the error increases, the estimated preference utilities are converging. Thus, applying the *weighted by probability* approach, which captures the difference between these utilities, might not be superior to the *threshold* approach in terms of lower uncertainty when choice error is high. This would be an additional argument for rigorously excluding inattentive respondents from consumer studies, delivering the basis for BiTURF (Jaeger et al., 2026b; Llobell et al., 2025).

Furthermore, another limitation pertains to the data sets used. Despite satisfying convergence and good fits when applying the respective HB models, both data sets are relatively sparse, potentially leading individual coefficients to shrink towards the population's mean (see, e.g., Orme & Chrzan, 2021, pp. 156–157; Train, 2009, pp. 266–267). However, regarding the MaxDiff data, the design allows for at least medium parameter precision (Lipovetsky et al., 2015), and the applied HB-MNL model has been shown to be robust against shrinkage even when having sparse data (see, e.g., Allenby et al., 2005; Lenk et al., 1996).

Regarding the CATA data, we only have one observation per item. However, using just one observation per coefficient is not uncommon, as it is, for example, the case in sparse MaxDiff (see, e.g., Chrzan & Peitz, 2019).

Finally, for the upper-level priors in the HB model, we assumed a single normal distribution as implemented in most proprietary software. We invite future researchers to test different distributions (e.g., a mixture of normals; Elshiewy et al., 2017) or different models (e.g., multivariate probit; Edwards & Allenby, 2003).

6 Conclusion

This study demonstrates how to apply Bayesian statistics to process input data for a TURF analysis and compares the results with those of a classical TURF analysis. It evaluates both the *threshold* and *weighted by probability* approaches to exemplary MaxDiff and CATA data. Generally, BiTURF analysis quantifies uncertainty in reach and frequency parameter estimates and provides researchers and practitioners with additional information to base their decisions on. Overall, the exemplary results of BiTURF and classical TURF analyses are quite similar. However, BiTURF analyses based on *weighted by probability* estimates tend to yield slightly more conservative reach and frequency estimates. In addition, this approach might also provide narrower uncertainty intervals than the *threshold* approach.

Statements and Declarations

CRediT authorship contribution statement

Joshua Benjamin Schramm: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Felix Josua Lang**: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization, Visualization. **Marcel Lichters**: Writing – review & editing, Software, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A step-by-step tutorial is published on the Open Science Framework (OSF Data Supplement to BiTURF: Quantifying Uncertainty to Enhance Strategic Decision-Making; Schramm et al., 2026). Both data sets are freely available and can be accessed via the corresponding publication.

References

Aizaki, H., & Fogarty, J. (2023). *R* packages and tutorial for case 1 best–worst scaling. *Journal of Choice Modelling*, *46*, Article 100394. https://doi.org/10.1016/j.jocm.2022.100394

Allenby, G. M., & Ginter, J. L. (1995). Using Extremes to Design Products and Segment Markets. *Journal of Marketing Research*, *32*(4), 392–403. https://doi.org/10.1177/002224379503200402

Allenby, G. M., Rossi, P. E., & McCulloch, R. E. (2005). Hierarchical Bayes Models: A Practitioners Guide. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.655541

Ares, G., Dauber, C., Fernández, E., Giménez, A., & Varela, P. (2014). Penalty analysis based on CATA questions to identify drivers of liking and directions for product reformulation. *Food Quality and Preference*, *32*, 65–76. https://doi.org/10.1016/j.foodqual.2013.05.014

Ares, G., & Jaeger, S. R. (2014). Check-all-that-apply (CATA) questions with consumers in practice: Experimental considerations and impact on outcome. In J. Delarue, J. B. Lawlor, & M. Rogeaux (Eds.), *Woodhead publishing series in food science, technology and nutrition: number 274. Rapid sensory profiling techniques and related methods: Applications in new product development and consumer research* (pp. 227–245). Woodhead. https://doi.org/10.1533/9781782422587.2.227

Ares, G., Varela, P., Rado, G., & Giménez, A. (2011). Identifying ideal products using three different consumer profiling methodologies. Comparison with external preference mapping. *Food Quality and Preference*, *22*(6), 581–591. https://doi.org/10.1016/j.foodqual.2011.04.004

Bi, J., & Kuesten, C. (2025). Product effect size estimation and performance accuracy (validity and reliability) assessment for CATA and TCATA data. *Food Quality and Preference*, *127*, Article 105441. https://doi.org/10.1016/j.foodqual.2025.105441

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

Chrzan, K., & Orme, B. K. (2019). *Applied MaxDiff: A Practitioner's Guide to Best-Worst Scaling*. Sawtooth Software.

Chrzan, K., & Peitz, M. (2019). Best-Worst Scaling with many items. *Journal of Choice Modelling*, *30*, 61–72. https://doi.org/10.1016/j.jocm.2019.01.002

Displayr. (2025). *TURF - Incrementality Plot*. https://wiki.q-researchsoftware.com/wiki/TURF_-_Incrementality_Plot

Edwards, Y. D., & Allenby, G. M. (2003). Multivariate Analysis of Multiple Response Data. *Journal of Marketing Research*, *40*(3), 321–334. https://doi.org/10.1509/jmkr.40.3.321.19233

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, *40*(1), 1–5. https://doi.org/10.1080/00031305.1986.10475342

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on statistics and applied probability: Vol. 57*. Chapman & Hall. https://doi.org/10.1201/9780429246593

Elshiewy, O., Guhl, D., & Boztuğ, Y. (2017). Multinomial Logit Models in Marketing – From Fundamentals to State-of-the-Art. *Marketing: ZFP – Journal of Research and Management*, *39*(3), 32–49. https://doi.org/10.15358/0344-1369-2017-3-32

Elshiewy, O., Zenetti, G., & Boztuğ, Y. (2016). Differences Between Classical and Bayesian Estimates for Mixed Logit Models: A Replication Study. *Journal of Applied Econometrics*, *32*(2), 470–476. https://doi.org/10.1002/jae.2513

Ennis, J. M., Fayle, C. M., & Ennis, D. M. (2012). eTURF: A competitive TURF algorithm for large datasets. *Food Quality and Preference*, *23*(1), 44–48. https://doi.org/10.1016/j.foodqual.2011.06.004

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, *6*, 733–807.

Gensler, S., Hinz, O., Skiera, B., & Theysohn, S. (2012). Willingness-to-pay estimation with choice-based conjoint analysis: Addressing extreme response behavior with individually adapted designs. *European Journal of Operational Research*, *219*(2), 368–378. https://doi.org/10.1016/j.ejor.2012.01.002

Hasted, A. (2023). Product Portfolio Management: Turf. In C. Gómez-Corona & H. Rodrigues (Eds.), *Methods and Protocols in Food Science. Consumer research methods in food science* (pp. 361–373). Humana Press. https://doi.org/10.1007/978-1-0716-3000-6_18

Hauser, J. R., Eggers, F., & Selove, M. (2019). The Strategic Implications of Scale in Choice-Based Conjoint Analysis. *Marketing Science*, *38*(6), 1059–1081. https://doi.org/10.1287/mksc.2019.1178

Howell, J. (2016). *A Simple Introduction to TURF Analysis.* Sawtooth Software, Inc. https://sawtoothsoftware.com/resources/technical-papers/a-simple-introduction-to-turf-analysis

Jaeger, S. R., Andersen, G. B. H., Chheang, S. L., & Llobell, F. (2026a). The ErrVarNorm (EVN) index in Case 1 Best–Worst Scaling: Further methodological and data quality investigations. *Food Quality and Preference*, *141*, Article 105892. https://doi.org/10.1016/j.foodqual.2026.105892

Jaeger, S. R., Beresford, M. K., Paisley, A. G., Antúnez, L., Vidal, L., Cadena, R. S., Giménez, A., & Ares, G. (2015). Check-all-that-apply (CATA) questions for sensory product characterization by consumers: Investigations into the number of terms used in CATA questions. *Food Quality and Preference*, *42*, 154–164. https://doi.org/10.1016/j.foodqual.2015.02.003

Jaeger, S. R., Chheang, S. L., Jin, D., Roigard, C. M., & Ares, G. (2020). Check-all-that-apply (CATA) questions: Sensory term citation frequency reflects rated term intensity and applicability. *Food Quality and Preference*, *86*, Article 103986. https://doi.org/10.1016/j.foodqual.2020.103986

Jaeger, S. R., Chheang, S. L., & Llobell, F. (2026b). Applications of the ErrVarNorm index in case 1 best-worst scaling and data quality insights. *Food Quality and Preference*, *136*, Article 105730. https://doi.org/10.1016/j.foodqual.2025.105730

Johnson, A. A., Ott, M. Q., & Dogucu, M. (2022). *Chapman & Hall/CRC texts in statistical science*. *Bayes Rules! An Introduction to Applied Bayesian Modeling*. Chapman and Hall/CRC.

Jürkenbeck, K., & Spiller, A. (2021). Importance of sensory quality signals in consumers' food choice. *Food Quality and Preference*, *90*, Article 104155. https://doi.org/10.1016/j.foodqual.2020.104155

Karvanen, J., Rantanen, A., & Luoma, L. (2014). Survey data and Bayesian analysis: A cost-efficient way to estimate customer equity. *Quantitative Marketing and Economics*, *12*(3), 305–329. https://doi.org/10.1007/s11129-014-9148-4

Klopčič, M., Slokan, P., & Erjavec, K. (2020). Consumer preference for nutrition and health claims: A multi-methodological approach. *Food Quality and Preference*, *82*, Article 103863. https://doi.org/10.1016/j.foodqual.2019.103863

Kuesten, C., & Bi, J. (2021). TURF analysis for CATA data using *R* package 'turfR'. *Food Quality and Preference*, *91*, Article 104201. https://doi.org/10.1016/j.foodqual.2021.104201

Lagerkvist, C. J., Okello, J., & Karanja, N. (2012). Anchored vs. relative best–worst scaling and latent class vs. hierarchical Bayesian analysis of best–worst choice data: Investigating the importance of food quality attributes in a developing country. *Food Quality and Preference*, *25*(1), 29–40. https://doi.org/10.1016/j.foodqual.2012.01.002

Lattery, K. (2010). Anchoring Maximum Difference Scaling Against a Threshold: Dual Response and Direct Binary Responses. In Sawtooth Software, Inc. (Ed.), *Proceedings of the Sawtooth Software Conference: October 2010* (pp. 91–106).

Lenk, P. J., Desarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science*, *15*(2), 173–191. https://doi.org/10.1287/mksc.15.2.173

Lichters, M., Möslein, R., Sarstedt, M., & Scharf, A. (2021). Segmenting Consumers Based on Sensory Acceptance Tests in Sensory Labs, Immersive Environments, and Natural Consumption Settings. *Food Quality and Preference*, *48*, Article 104138. https://doi.org/10.1016/j.foodqual.2020.104138

Lipovetsky, S., Liakhovitski, D., & Conklin, M. (2015). What is the Right Size for my MaxDiff Study? In Sawtooth Software, Inc. (Chair), *Sawtooth Software Conference.* Symposium conducted at the meeting of Sawtooth Software, Inc., Orlando, FL.

Llobell, F., Choisy, P., Chheang, S. L., & Jaeger, S. R. (2025). Measurement and evaluation of participant response consistency in Case 1 Best-Worst-Scaling (BWS) in food consumer science. *Food Quality and Preference*, *123*, Article 105335. https://doi.org/10.1016/j.foodqual.2024.105335

Loaiza-Maya, R., & Nibbering, D. (2022). *Fast variational Bayes methods for multinomial probit models.* https://doi.org/10.48550/arxiv.2202.12495

Louviere, J., Lings, I., Islam, T., Gudergan, S., & Flynn, T. (2013). An introduction to the application of (case 1) best–worst scaling in marketing research. *International Journal of Research in Marketing*, *30*(3), 292–303. https://doi.org/10.1016/j.ijresmar.2012.10.002

Mersmann, O., Beleites Clauida, Hurling, R., Friedman, A., & Ulrich, J. M. (2024). *microbenchmark: Accurate Timing Functions* (Version 1.5.0) [Computer software]. CRAN. https://CRAN.R-project.org/package=microbenchmark

Miaoulis, G., Parsons, H., & Free, V. (1990). Turf: A New Planning Approach for Product Line Extensions. *Marketing Research*, *2*(1), 28–40.

Naughton, P., Schramm, J. B., & Lichters, M. (2025). The eyes eat first: Improving consumer acceptance of plant-based meat alternatives by adjusting front-of-pack labeling. *Food Quality and Preference*, *131*, Article 105567. https://doi.org/10.1016/j.foodqual.2025.105567

Orme, B. K. (2000). *Hierarchical Bayes: Why All the Attention?* Sawtooth Software, Inc. Sawtooth Software Research Paper Series. https://sawtoothsoftware.com/resources/technical-papers/hierarchical-bayes-why-all-the-attention

Orme, B. K. (2009). *Anchored Scaling in MaxDiff Using Dual Response.* Sawtooth Software, Inc. https://sawtoothsoftware.com/resources/technical-papers/anchored-scaling-in-maxdiff-using-deal-response

Orme, B. K. (2019). *Consistency Cutoffs to Identify "Bad" Respondents in CBC, ACBC, and MaxDiff.* Sawtooth Software, Inc. Sawtooth Software Research Paper Series. https://sawtoothsoftware.com/resources/technical-papers/consistency-cutoffs-to-identify-bad-respondents-in-cbc-acbc-and-maxdiff

Orme, B. K. (2021). *The CBC/HB System: Technical Paper V5.6.* Sawtooth Software, Inc. Sawtooth Software Technical Paper Series. https://sawtoothsoftware.com/resources/technical-papers/cbc-hb-technical-paper

Orme, B. K., & Chrzan, K. (2021). *Becoming an Expert in Conjoint Analysis: Choice Modeling for Pros* (2nd ed.). Sawtooth Software, Inc.

Payne, J., Talavera, M., & Koppel, K. (2023). Consumer perceptions of hotel shampoos and lotions. *Journal of Sensory Studies*, *38*(3), Article e12817. https://doi.org/10.1111/joss.12817

Price, S., Viglia, G., Hartwell, H., Hemingway, A., Chapleo, C., Appleton, K., Saulais, L., Mavridis, I., & Perez-Cueto, F. J. (2016). What are we eating? Consumer information requirement within a workplace canteen. *Food Quality and Preference*, *53*, 39–46. https://doi.org/10.1016/j.foodqual.2016.05.014

quantilope. (2024). *What Is TURF Analysis? Examples and How To Use in Market Research*. https://www.quantilope.com/resources/what-is-turf-analysis-and-how-to-use-it-in-your-customer-research

Rooderkerk, R. P., & Lehmann, D. R. (2021). Incorporating Consumer Product Categorizations into Shelf Layout Design. *Journal of Marketing Research*, *58*(1), 50–73. https://doi.org/10.1177/0022243720964127

Rooderkerk, R. P., van Heerde, H. J., & Bijmolt, T. H. A. (2013). Optimizing Retail Assortments. *Marketing Science*, *32*(5), 699–715. https://doi.org/10.1287/mksc.2013.0800

Rossi, P. E., Allenby, G. M., & Misra, S. (2024). *Bayesian Statistics and Marketing* (2nd ed.). *Wiley series in probability and statistics*. Wiley.

Sablotny-Wackershauser, V., Lichters, M., Guhl, D., Bengart, P., & Vogt, B. (2024). Crossing incentive alignment and adaptive designs in choice-based conjoint: A fruitful endeavor. *Journal of the Academy of Marketing Science*, *52*(3), 610–633. https://doi.org/10.1007/s11747-023-00997-5

Sawtooth Software, Inc. (2025a). *Lighthouse Studio 9* (Version 9.16.16) [Computer software]. Sawtooth Software Inc. https://www.sawtoothsoftware.com/products/online-surveys

Sawtooth Software, Inc. (2025b). *TURF*. https://sawtoothsoftware.com/help/maxdiff-analyzer/manual/index.html?turf.html

Schramm, J. B., Lang, F. J., & Lichters, M. (2026). *OSF Data Supplement to BiTURF: Quantifying Uncertainty to Enhance Strategic Decision-Making*. https://doi.org/10.17605/OSF.IO/FDH24

Schramm, J. B., & Lichters, M. (2025a). Incentive alignment in anchored MaxDiff yields superior predictive validity. *Marketing Letters*, *36*, 1–16. https://doi.org/10.1007/s11002-023-09714-2

Schramm, J. B., & Lichters, M. (2025b). validateHOT - an *R* package for the analysis of holdout/validation tasks and other choice modeling tools. *Journal of Open Source Software*, *10*(107), Article 6708. https://doi.org/10.21105/joss.06708

Schramm, J. B., Lichters, M., & Guhl, D. (2025). *Anchoring Methods for MaxDiff and Their Predictive Validity in Product Choice* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.5674983

Schreiner, T., & Baier, D. (2021). Online retailing during the COVID-19 pandemic: consumer preferences for marketing actions with consumer self-benefits versus other-benefit components. *Journal of Marketing Management*, *37*(17-18), 1866–1902. https://doi.org/10.1080/0267257X.2022.2030784

Schuster, A. L., Crossnohere, N. L., Campoamor, N. B., Hollin, I. L., & Bridges, J. F. (2024). The rise of best-worst scaling for prioritization: A transdisciplinary literature review. *Journal of Choice Modelling*, *50*, Article 100466. https://doi.org/10.1016/j.jocm.2023.100466

Serra, D. (2013). Implementing TURF analysis through binary linear programming. *Food Quality and Preference*, *28*(1), 382–388. https://doi.org/10.1016/j.foodqual.2012.10.001

Stan Development Team. (2025). *RStan: the R interface to Stan* (Version 2.32.7) [Computer software]. https://mc-stan.org/

Train, K. E. (2009). *Discrete Choice Methods with Simulation* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511805271

Veenman, M., Stefan, A. M., & Haaf, J. M. (2024). Bayesian hierarchical modeling: An introduction and reassessment. *Behavior Research Methods*, *56*(5), 4600–4631. https://doi.org/10.3758/s13428-023-02204-3

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Appendix

1 Illustration of the coding of the design matrix for anchored MaxDiff variants

Assuming we conduct a MaxDiff study (i.e., Best-Worst Scaling Case 1) with the following settings: $K = 7$ (the number of items in the MaxDiff; seven different Pizza alternatives displayed in Table A1), $T = 7$ (the number of tasks per individual), $J = 3$ (the number of item alternatives per task). One of the MaxDiff tasks participant *n* faces is displayed in Fig. A1.

Table A1
Seven pizza alternatives considered for the MaxDiff example.

No.	Pizza
1	Margherita
2	Ham
3	Salami
4	Vegetarian
5	Hawaii
6	Funghi
7	Tonno

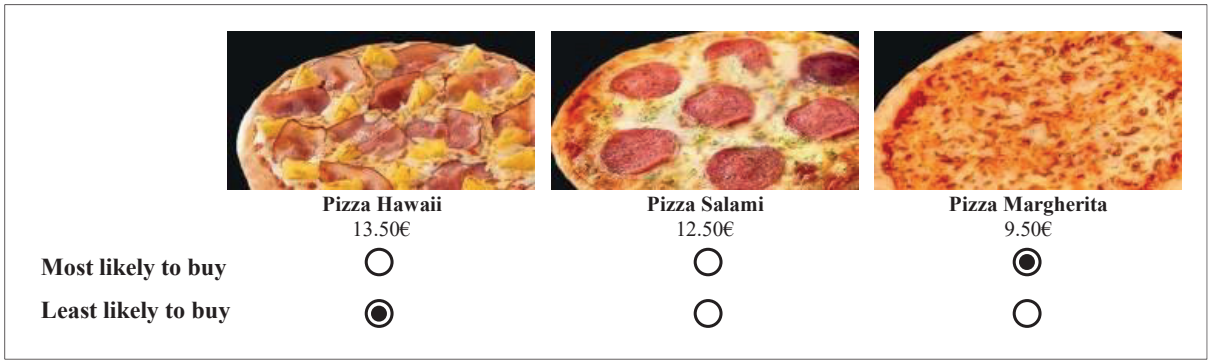

Fig. A1. Exemplary MaxDiff task based on stimuli by Sablotny-Wackershauser et al. (2024).

Participant *n* chooses *Pizza Margherita* as the alternative that s/he is most likely to buy, while *Pizza Hawaii* is the alternative that s/he is least likely to buy. We apply the best-worst coding approach (also known as the marginal coding approach; Aizaki & Fogarty, 2023). Applying the best-worst coding approach results in the design matrix in Table A2.

Table A2
Illustration of the best-worst coding of the MaxDiff choice in Fig. 1.

id	cs	alt	margherita	ham	salami	vegetarian	hawaii	funghi	choice
1	1	1	1	0	0	0	0	0	1
1	1	2	0	0	1	0	0	0	0
1	1	3	0	0	0	0	1	0	0
1	2	1	-1	0	0	0	0	0	0
1	2	2	0	0	-1	0	0	0	0
1	2	3	0	0	0	0	-1	0	1
…									

Table A3
Illustration of the coding of the direct anchor questions in Fig. 2.

id	alt	margherita	ham	salami	vegetarian	hawaii	funghi	tonno	choice
1	1	1	0	0	0	0	0	0	1
1	2	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	0	0	0
1	2	0	0	0	0	0	0	0	1
…									

Here, the seventh alternative (*Pizza Tonno*) is not coded into a separate column to prevent linear dependency (Chrzan & Orme, 2019, p. 21). The best ("most likely to buy") and worst ("least likely to buy") coded items are identical, except that the worst alternatives are coded as '-1'. The participant's best *Pizza Margherita* and worst choice *Pizza Hawaii* are assigned a '1' in the choice column.

The results of the preference estimation are relative utilities (Lagerkvist et al., 2012). To make these utilities absolute, one can apply anchoring. In the manuscript, we used the direct anchor approach proposed by Lattery (2010). When using a direct anchor, participants see all or a subset of the items again and must state whether each item is a purchase option or not (Schramm & Lichters, 2025). The way the anchor is implemented can vary. While, for example, Schramm and Lichters (2025) used a binary forced-choice in which participants had to state for each alternative whether it was a purchase option (or not), Lattery (2010) implemented the direct anchor via a multiple select list. An example of the direct anchor, as implemented in the paper by Schramm and Lichters (2025), with exemplary responses, is displayed in Fig. A2.

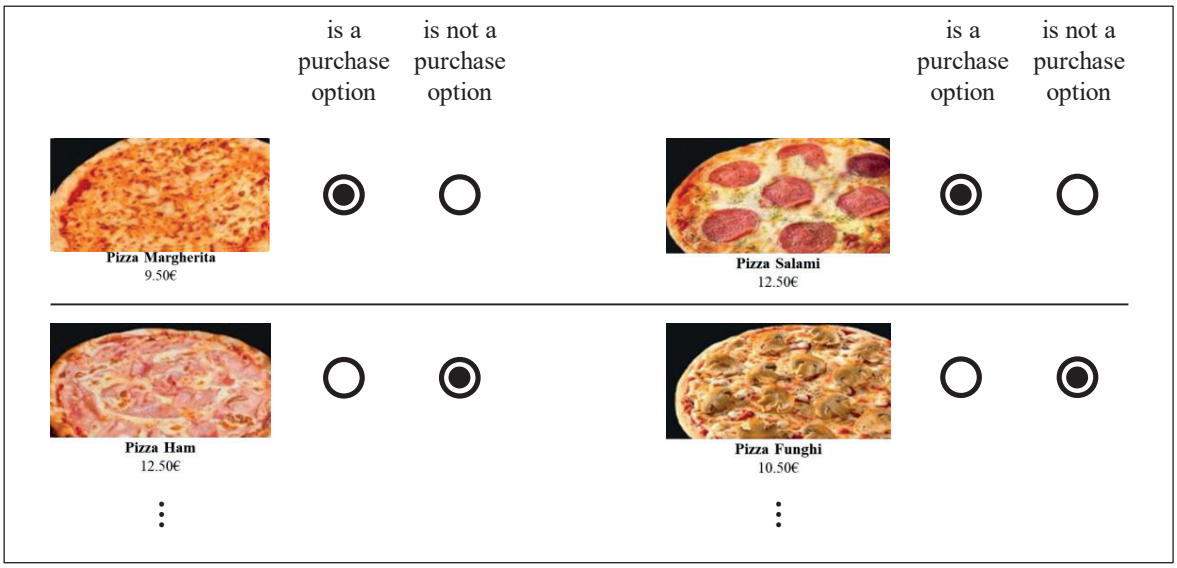

Fig. A2. Exemplary direct anchor question based on stimuli by Sablotny-Wackershauser et al. (2024).

When implementing the direct anchor, the MaxDiff coding is similar to that in Table A3. The only difference from the best-worst coding shown in Table A2 is that the seventh alternative (*Pizza Tonno*) is now also displayed, since the anchor is omitted during estimation to prevent linear dependency (Chrzan & Orme, 2019, p. 21).

Let us only focus on the best choice (i.e., *Pizza Margherita*). Here, two new rows are added to the design matrix. The first describes the alternative for which the item is selected as a purchase option (i.e., this is *Pizza Margherita*), and the second describes the anchor (i.e., later transformed into the no-buy utility). Here, the participant stated that the *Pizza Margherita* is a purchase option (see Fig. A2), thus, the row describing *Pizza Margherita* (*alt = 1*) is assigned a '1' in the choice column. For *Pizza Ham* the participant stated that it is not a purchase option. Thus, in this scenario, the row describing the anchor (i.e., the no-buy alternative, *alt = 2*) is assigned a '1'. We refer interested readers who would like to learn more about coding approaches of the MaxDiff and anchor tasks to the following references: Aizaki & Fogarty, 2023; Chrzan & Orme, 2019; Lattery, 2010; Schramm & Lichters, 2025. For other anchor approaches for anchoring MaxDiff, we refer readers to the paper by Schramm et al. (2025).

2 Summary statistics and diagnostics for the Bayes models for MaxDiff and CATA

Table A4
Summary statistics for the hierarchical Bayes multinomial logit model applied to the MaxDiff data

	Upper-level draws (b)			Lower-level draws (β)	
	Raw utilities	$\hat{R}$	Effective sample size	Raw utilities	Choice probabilities
G1	-1.59 [-2.85, -0.35]	1.00	2771	-1.63 [-2.12, -1.16]	41.71 [39.36, 44.12]
G2	-0.90 [-1.63, -0.16]	1.00	2450	-0.90 [-1.22, -0.58]	43.03 [39.86, 46.28]
G3	-0.60 [-1.40, 0.19]	1.00	2725	-0.59 [-0.91, -0.27]	41.36 [38.53, 44.28]
G4	-1.06 [-1.82, -0.25]	1.00	2853	-1.07 [-1.41, -0.74]	40.62 [37.63, 43.67]
G5	0.63 [-0.35, 1.65]	1.00	2715	0.64 [0.26, 1.03]	56.05 [53.73, 58.41]
G6	0.44 [-0.34, 1.24]	1.00	2764	0.44 [0.12, 0.77]	53.84 [50.75, 57.06]
G7	-1.62 [-2.41, -0.85]	1.00	3012	-1.63 [-1.98, -1.29]	34.62 [31.82, 37.53]
G8	-2.21 [-3.00, -1.43]	1.00	2994	-2.24 [-2.62, -1.87]	29.49 [26.91, 32.05]
G9	-4.29 [-5.53, -3.12]	1.00	3031	-4.37 [-4.98, -3.82]	24.13 [21.79, 26.53]
G10	0.05 [-0.80, 0.90]	1.00	2442	0.04 [-0.28, 0.37]	51.93 [48.99, 54.85]
G11	0.20 [-0.63, 1.05]	1.00	2865	0.20 [-0.15, 0.54]	53.26 [50.54, 56.03]
G12	-0.52 [-1.18, 0.11]	1.00	2611	-0.52 [-0.82, -0.22]	44.02 [40.76, 47.18]
G13	-0.10 [-0.72, 0.52]	1.00	2919	-0.11 [-0.42, 0.20]	48.34 [44.75, 51.83]
G14	-3.36 [-4.19, -2.53]	1.00	2437	-3.40 [-3.84, -2.98]	21.48 [18.72, 24.29]
G15	0.61 [-0.23, 1.43]	1.00	2929	0.59 [0.25, 0.95]	55.57 [52.59, 58.57]
G16	-4.16 [-5.06, -3.28]	1.00	2978	-4.23 [-4.71, -3.76]	19.30 [16.97, 21.74]

Note: The anchor parameter was fixed to 0 to prevent linear dependency during the estimation | Root likelihood across all iterations = 0.71 [0.70, 0.72] | Null log-likelihood = -6210.60 | Model's log-likelihood = -1815.75 (-1932.42 [-2003.04, -1861.93]) | Log marginal density = -2071.45 | predictive accuracy = 0.09 [0.07, 0.11]. The measure compares how well the posterior draws predict the actual choice data vs. chance level (< .001) based on a manual calculation (see the *R* tutorial for details).

Table A5
Summary statistics for the hierarchical Bayes logit model applied to the CATA data

	Upper-level draws (b)			Lower-level draws (β)	
	Raw utilities	$\hat{R}$	Effective sample size	Raw utilities	Choice probabilities
S1	-0.26	1.00	2590	-0.25	47.50
	[-1.03, 0.39]			[-0.81, 0.24]	[42.91, 52.33]
S2	-3.39	1.00	2178	-3.38	17.60
	[-5.58, -2.03]			[-5.49, -2.08]	[13.95, 21.51]
S3	0.38	1.00	2769	0.39	54.52
	[-0.26, 1.15]			[-0.07, 0.99]	[49.78, 59.37]
S4	-0.83	1.00	2592	-0.82	41.07
	[-1.76, -0.20]			[-1.58, -0.29]	[36.56, 45.92]
S5	-4.39	1.00	2216	-4.39	12.12
	[-6.97, -2.68]			[-6.98, -2.71]	[9.12, 15.68]
S6	-1.15	1.00	2069	-1.14	37.46
	[-2.23, -0.45]			[-2.06, -0.51]	[32.89, 42.08]
S7	-0.64	1.00	2517	-0.63	43.96
	[-1.54, 0.05]			[-1.33, -0.09]	[39.50, 48.48]
S8	-5.37	1.00	2738	-5.39	6.64
	[-8.74, -3.04]			[-8.73, -3.06]	[4.05, 9.46]
S9	-6.05	1.00	2747	-6.07	11.60
	[-8.88, -3.88]			[-8.75, -3.96]	[9.00, 14.32]

Note: The anchor parameter was fixed to 0 to prevent linear dependency during the estimation | Root likelihood across all iterations = 0.78 [0.75, 0.81] | Null log-likelihood = -1247.67 | Model's log-likelihood = -338.43 (-471.65 [-553.03, -394.79]) | Log marginal density = -597.83 | predictive accuracy = 0.20 [0.14, 0.27]. The measure compares how well the posterior draws predict the actual choice data vs. chance level (= .002) based on a manual calculation (see the *R* tutorial for details).

References

Aizaki, H., & Fogarty, J. (2023). *R* packages and tutorial for case 1 best–worst scaling. *Journal of Choice Modelling*, *46*, Article 100394. https://doi.org/10.1016/j.jocm.2022.100394

Chrzan, K., & Orme, B. K. (2019). *Applied MaxDiff: A Practitioner's Guide to Best-Worst Scaling*. Sawtooth Software.

Lagerkvist, C. J., Okello, J., & Karanja, N. (2012). Anchored vs. relative best–worst scaling and latent class vs. hierarchical Bayesian analysis of best–worst choice data: Investigating the importance of food quality attributes in a developing country. *Food Quality and Preference*, *25*(1), 29–40. https://doi.org/10.1016/j.foodqual.2012.01.002

Lattery, K. (2010). Anchoring Maximum Difference Scaling Against a Threshold: Dual Response and Direct Binary Responses. In Sawtooth Software, Inc. (Ed.), *Proceedings of the Sawtooth Software Conference: October 2010* (pp. 91–106).

Sablotny-Wackershauser, V., Lichters, M., Guhl, D., Bengart, P., & Vogt, B. (2024). Crossing incentive alignment and adaptive designs in choice-based conjoint: A fruitful endeavor. *Journal of the Academy of Marketing Science*, *52*(3), 610–633. https://doi.org/10.1007/s11747-023-00997-5

Schramm, J. B., & Lichters, M. (2025). Incentive alignment in anchored MaxDiff yields superior predictive validity. *Marketing Letters*, *36*, 1–16. https://doi.org/10.1007/s11002-023-09714-2

Schramm, J. B., Lichters, M., & Guhl, D. (2025). *Anchoring Methods for MaxDiff and Their Predictive Validity in Product Choice* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.5674983

Otto von Guericke University Magdeburg
Faculty of Economics and Management
P.O. Box 4120 | 39016 Magdeburg | Germany

Tel.: +49 (0) 3 91 / 67-1 85 84
Fax: +49 (0) 3 91 / 67-1 21 20

www.fww.ovgu.de/femm

ISSN 1615-4274