Augmenting Conventional Evaluation Charts with MCMC-Based Bayesian Probabilities

Tomoyasu Nakano [1] and Masataka Goto[1]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan

Abstract

Evaluations in information visualization and HCI research frequently compare group means using bar charts with error bars, box plots, violin plots, or paired plots, often marking differences with asterisks derived from p-values and fixed significance levels. However, figures grounded in null-hypothesis significance testing (NHST) provide limited insight into practical differences, such as the probability that an effect is practically meaningful and how this probability varies with the choice of threshold. We define a practically meaningful hypothesis U about a posterior quantity (e.g., $U \equiv \mu_i - \mu_j > c$) and compute the Bayesian Probability that the Hypothesis is Correct (BPHC) directly from Markov chain Monte Carlo (MCMC) generated quantities. We then augment conventional charts by replacing binary significance markers with BPHC visualizations that relate effect-size thresholds to posterior probabilities. We present scalar BPHC values at specified thresholds, as well as one- and two-dimensional BPHC visualizations over thresholds for richer contexts. For conjunctive hypotheses, matched MCMC samples are summarized using paired plots with discrete coloring to indicate joint exceedance. We extend familiar evaluation charts into a Bayesian setting to lower barriers to use and understanding, offering this work in progress to foster discussion.

CCS Concepts

• ***Mathematics of computing*** → *Statistical graphics;*

1. Introduction

In Information Visualization (InfoVis) and Human-Computer Interaction (HCI), researchers frequently compare task performance, accuracy (error rates), or subjective ratings across multiple conditions [SSV*22]. These measurements are typically collected as quantitative numerical data, enabling the analysis of differences between distributions and statistical summaries such as means and variances. To visualize such differences, researchers commonly use bar charts with error bars, box plots, or violin plots. These comparisons are typically embedded within a null-hypothesis significance testing (NHST) framework, where a no-difference null hypothesis is tested and figures are annotated with asterisks based on *p*-values and significance levels (Fig. 1, first column). Although reporting effect sizes alongside *p*-values has been recommended, asterisk-based NHST annotations convey only null-hypothesis rejection and provide little insight into practical significance or uncertainty.

While various limitations of NHST have been discussed [AGM19, WSL19, Toy24a], we argue for complementing NHST with Bayesian results where appropriate. Bayesian statistics offer advantages [Kru10a, Kru10b, KR12, KNH16, Toy24a, Toy24b] that are difficult to achieve through NHST alone. However, the practical use and interpretation of Bayesian results remain relatively challenging. Accordingly, we extend conventional evaluation charts to present Bayesian results in a familiar yet more informative way.

2. Augmenting Evaluation Charts with Bayesian Results

To extend evaluation charts, we introduce Toyoda's "Probability that the Hypothesis is Correct (PHC)" [Toy24a, Toy24b], but to avoid confusion we refer to it here as the *Bayesian* Probability that the Hypothesis is Correct (BPHC). This does *not* denote the probability that a scientific hypothesis is absolutely "true"; rather, BPHC represents a degree of belief in the Bayesian sense. A hypothesis U can be, for example, $U \equiv (\mu_j - \mu_i) > c$. To compute BPHC, we sample from the posterior distribution using Markov chain Monte Carlo (MCMC) with the No-U-Turn Sampler (NUTS) [HG14], and estimate BPHC from the resulting generated quantities. In this section, we describe how evaluation charts can be augmented using BPHC, covering scalar, one-dimensional, two-dimensional, and parallel-coordinates cases. While these visualizations build on familiar chart forms, their integration into a Bayesian framework lowers the barrier to adoption and improves interpretability.

2.1. MCMC-based generated quantities and BPHC

MCMC-based generated quantities $g(\theta^{(t)})$, defined as functions of samples $\theta^{(t)}$ drawn via MCMC, provide a basis for probabilistic inference by approximating the posterior distribution and enabling estimation of expectations, credible intervals, and other summaries. Treating these samples and their generated quantities as data allows the direct use of conventional visualization methods. For ex-

DOI: 10.2312/evs.20261007

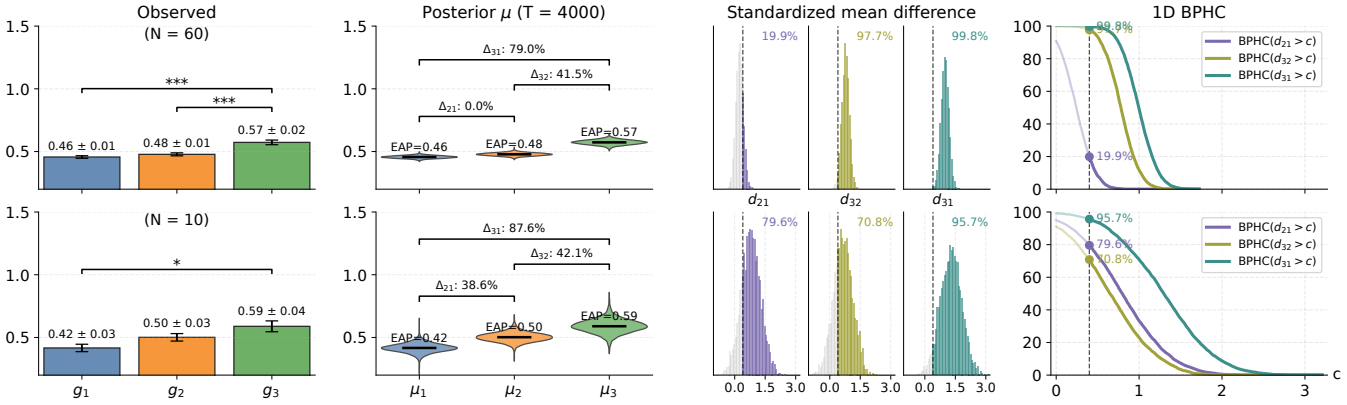

Figure 1: *Examples of synthetic observed-data summaries and MCMC-based Bayesian visualizations illustrating scalar and 1D BPHC. From left to right: bar charts with standard-error bars and NHST annotations (two-sided Welch's t-tests with Holm correction; $^{*}p < 0.05$, $^{***}p < 0.001$); posterior means μ_i with scalar BPHC for $\Delta_{ij} > 0.1$; posterior standardized mean differences with scalar BPHC for $d_{ij} > 0.4$; and 1D BPHC. Rows correspond to $N = 60$ and $N = 10$ per group. Synthetic data were generated from normal distributions with true means $(0.45, 0.50, 0.58)$ and standard deviations $(0.10, 0.12, 0.14)$. $N = 10$ uses the first 10 samples per group, leading to increased uncertainty and smoother 1D BPHC. Weakly informative priors were used for inference: $\mu_i \sim N(0,5)$ and $\sigma_i \sim$ Half-Student-t$(3,0,2)$.*

ample, Kruschke proposed BEST (Bayesian Estimation) [Kru13] as an alternative to the t test, interpreting posterior distributions using interval-based summaries such as regions of practical equivalence (ROPEs). Along similar lines, hypotheses can be defined as functions of posterior generated quantities, allowing the probability that they exceed (or fall below) a reference point (threshold) c to be computed explicitly. In psychology, Toyoda introduced this probability as PHC (which we refer to as BPHC) to provide a more intuitive way to communicate statistical evidence than traditional significance testing [Toy24a, Toy24b].

Let $\Delta_{ij} = \mu_i - \mu_j$ and let $c_{ij} > 0$ denote a domain-meaningful threshold. We also define the standardized mean difference $d_{ij} = (\mu_i - \mu_j)/\sqrt{(\sigma_i^2 + \sigma_j^2)/2}$. For each MCMC draw t, let $U^{(t)}$ denote the logical indicator of whether the hypothesis U holds for the t-th posterior sample. The indicator function $\mathbb{I}(U^{(t)})$ returns 1 if $U^{(t)}$ holds at iteration t, and 0 otherwise. The resulting probability is

$$\mathrm{BPHC}(U \equiv \Delta_{ij} > c_{ij}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left(\Delta_{ij}^{(t)} > c_{ij}\right). \tag{1}$$

For a set of index pairs J and thresholds $\{c_{ij}\}_{(ij)\in J}$, estimate the conjunctive probability by

$$\mathrm{BPHC}\left(U \equiv \bigwedge_{(ij)\in J} (\Delta_{ij} > c_{ij})\right) = \frac{1}{T}\sum_{t=1}^{T} \prod_{(ij)\in J} \mathbb{I}(\Delta_{ij}^{(t)} > c_{ij}). \tag{2}$$

Such probabilities have also been applied in cognitive science. Ogata *et al.* [O*23] analyzed the results of a cognitive psychology experiment on the Bouba–Kiki effect using Bayesian inference with MCMC. They visualized the posterior distributions of pairwise mean differences and reported the probabilities of hypotheses by setting the thresholds $c_{\cdot} = 0$ in Equations (1) and (2).

2.2. Scalar BPHC

Instead of plotting observed data distributions, we visualize posterior distributions of group means and explicitly report, for each pair of groups, the posterior probability that their means differ (Fig. 1, second column). To avoid the well-known misinterpretation of error bars [CG14, SSKQ25], we use violin plots. The key contribution is replacing NHST asterisks with explicit threshold-based posterior probabilities.

For two-group comparisons, the difference in means can be represented as a single posterior distribution using MCMC-generated quantities (Fig. 1, third column), an approach advocated as an alternative to the t test [Kru13, Toy24b] and adopted in applied evaluations [O*23]. By converting this distribution into exceedance probabilities, the degree of confidence becomes explicit. Following prior work, we encode these probabilities directly within the distribution using threshold-based coloring illustrated in the online documentation of the ggdist package [Kay24]. Here, differences are shown on the horizontal axis to align with the 1D BPHC (fourth column), whereas showing differences on the vertical axis would better align with the posterior distributions in the second column.

2.3. 1D BPHC

While scalar BPHC supports probabilistic interpretation at a fixed threshold, it becomes less practical when no non-arbitrary threshold can be specified. To address this, Toyoda proposed the PHC curve [Toy24a, Toy24b], which visualizes posterior probabilities across thresholds. We refer to this representation as the 1D BPHC.

We also apply threshold-based coloring to the 1D BPHC (Fig. 1, fourth column), ensuring visual consistency with the color-encoded posterior distributions (third column). Using a consistent threshold-based coloring scheme, effect size, uncertainty, and the degree of confidence can be conveyed simultaneously, supporting a coherent Bayesian interpretation across representations.

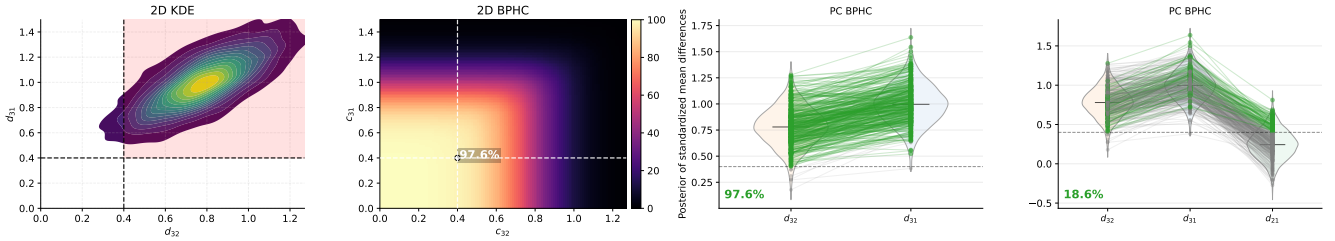

Figure 2: *Joint posterior visualizations for the $N = 60$ example in the top row of Fig. 1, based on standardized mean differences. From left to right: a 2D KDE and 2D BPHC for (d_{32}, d_{31}), followed by PC BPHC for the conjunctions of (d_{32}, d_{31}) and (d_{32}, d_{31}, d_{21}). Green lines indicate MCMC iterations where all conjunctive threshold conditions hold jointly.*

2.4. 2D BPHC

For conjunction hypotheses involving multiple pairwise comparisons, results are often examined through separate posterior distributions [O*23]. Instead, the joint distribution of posterior differences can be visualized directly. Figure 2 (first column) shows a two-dimensional kernel density estimate (KDE) of two posterior differences. Using the example with $N = 60$ shown in the top row of Fig. 1, the visualization represents the posterior probabilities that g_3 exceeds both g_2 and g_1, based on standardized mean differences d_{ij}. As in scalar and 1D BPHC, threshold-based coloring highlights the region where both differences exceed the specified threshold.

Summarizing the same paired MCMC samples yields a 2D BPHC (Fig. 2, second column), which directly reports the posterior probability that both differences hold simultaneously, enabling probabilistic interpretation of conjunction hypotheses that is difficult to obtain from marginal distributions alone.

2.5. Parallel-Coordinates BPHC (PC BPHC)

While 2D BPHC supports conjunctions of two pairwise differences, many analyses require reasoning over three or more related comparisons, such as ordering among multiple groups [O*23]. To support such higher-dimensional conjunction hypotheses, we introduce Parallel-Coordinates BPHC (PC BPHC). PC BPHC visualizes the posterior probability that conjunctive hypotheses hold using a parallel-coordinates representation. By combining violin plots with parallel coordinates, PC BPHC shows individual distributions and represents the probability that a hypothesis holds using polylines. MCMC-generated quantities from the same iteration are plotted across axes and connected with line segments, where each polyline corresponds to a single MCMC draw evaluated across all hypotheses. Color and/or opacity indicate whether a given draw exceeds the specified thresholds on all axes, and the proportion of highlighted lines reflects the joint posterior probability.

Figure 2 illustrates PC BPHC for conjunctions of two hypotheses (third column) and three hypotheses (fourth column). While PC BPHC can also be applied to conjunctions of two hypotheses, its advantage becomes most apparent when reasoning about three or more conditions. This visualization assumes a fixed threshold configuration and is therefore appropriate when meaningful thresholds can be discussed a priori. In practice, MCMC sampling typically produces a large number of draws (e.g., 4,000 paired samples in default Stan settings), and plotting all samples would result in visual clutter. To mitigate this, we apply uniform subsampling and visualize 300 draws, which provides an interpretable approximation of the posterior structure. Investigating alternative subsampling or aggregation strategies, and how uniform subsampling compares to them, is an important direction for future work. The results show that g_3 exceeds the other groups with high probability (97.6%), while the conjunctive hypothesis that higher category IDs consistently yield higher values (i.e., $g_1 < g_2$, $g_2 < g_3$, and $g_1 < g_3$) is not supported (18.6%). PC BPHC thus enables reasoning about higher-order conjunctions beyond scalar, 1D, or 2D BPHC.

3. Example of Actual Analysis

Until the preceding sections, our discussion was based on simulations using synthetic data and mean parameters. In this section, we apply the proposed visualization methods to the results of our previous work [NG24], which examined singing-skill annotations provided by multiple human evaluators for vocal performances. The dataset comprises 140 Japanese-language singing recordings.

The database includes 20 original songs from the RWC Music Database (Popular Music) [GHNO02], along with 120 cover recordings in which six additional singers performed each song. Of the 20 original songs, 10 were sung by male singers and the remaining 10 by female singers. The 120 cover recordings were performed by 40 singers (20 male and 20 female) with diverse singing experience, each singing three songs. All recordings were evaluated by 10 experts (five male and five female) specializing in music and vocal performance. The evaluations were conducted on mixed audio tracks containing both vocals and accompaniment, using six criteria—pitch, rhythm, pronunciation, expression, vocal projection, and overall performance—on a seven-point Likert scale. To ensure consistency across annotators, we provided a reference standard for singing skill and presented concrete singing examples corresponding to each of the seven scale points prior to annotation.

3.1. Visualization Results

We employed the graded response model (GRM) [Sam69] to aggregate each vocal performance's singing skill into a latent trait θ, while simultaneously estimating evaluator characteristics using

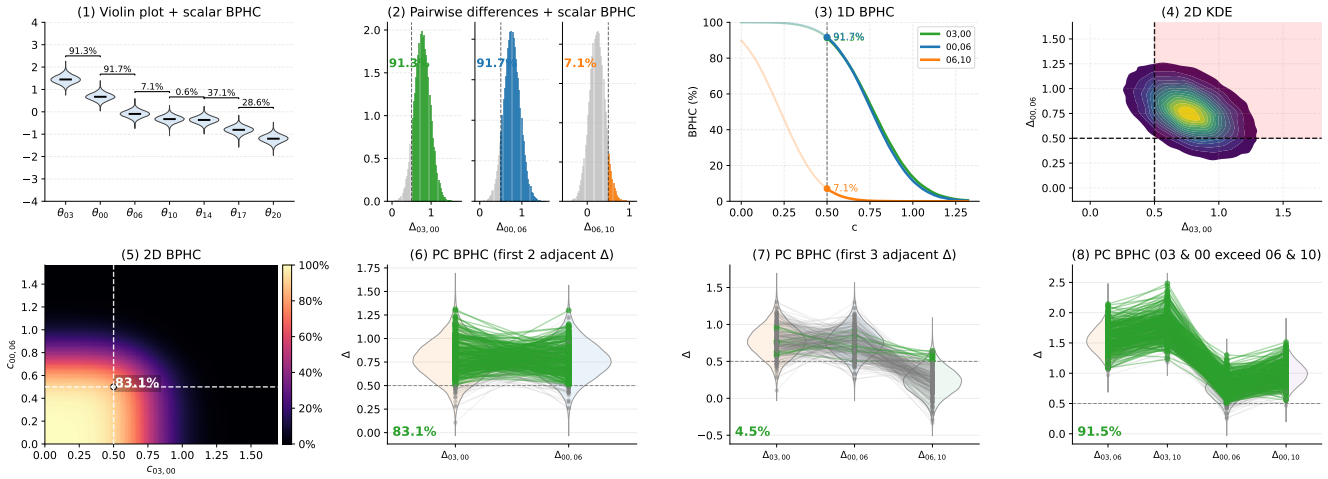

Figure 3: *Posterior distributions of singing skill θ for seven singers (03, 00, 06, 10, 14, 17, and 20) who performed the same song in RWC-MDB-P-2001 No. 12, visualized as violin plots. (1) Violin plots of the posterior distributions of singing skill θ, where horizontal brackets indicate adjacent pairwise comparisons with labels reporting* $\mathrm{BPHC}(\Delta_{ij} > 0.5)$. *(2, 3) Posterior distributions of pairwise differences and their corresponding 1D BPHC across thresholds. (4) A 2D KDE of two pairwise differences, highlighting the joint exceedance region. (5) The corresponding 2D BPHC. (6)–(8) PC BPHC, where green lines indicate posterior draws in which two, three, or four hypotheses are satisfied simultaneously.*

MCMC. This procedure yields the posterior distribution of θ for each of the 140 performances. Figure 3 presents an example for one song performed by seven singers. Because a standard normal prior $N(0,1)$ was used for θ, the posterior samples typically lie within approximately $[-3,3]$. Accordingly, we set the reference threshold to $c = 0.5$, corresponding to a difference of half a standard deviation, and report the resulting BPHC values throughout Fig. 3.

Figure 3 summarizes these results. Figure 3(1) shows the posterior distributions of singing skill θ for the seven singers, together with adjacent pairwise comparisons annotated by $\mathrm{BPHC}(\Delta_{i,j} = (\theta_i - \theta_j) > 0.5)$. Pairwise differences $\Delta_{03,00}$ and $\Delta_{00,06}$ exceed the threshold with high probability (91.3% and 91.7%, respectively; Fig. 3(1–3)), while their conjunction holds with probability 83.1% (Fig. 3(5, 6)). Figure 3(4) further illustrates the joint posterior distribution of these two pairwise differences, highlighting the region in which both exceed the threshold simultaneously. Accordingly, the ordering of the top three singers is supported (83.1%; Fig. 3(6)), whereas extending this constraint to the top four is much less plausible (4.5%; Fig. 3(7)). In addition, singers 00 and 03 are likely to consistently exceed singers 06 and 10 (91.5%; Fig. 3(8)).

4. Related Work

Visualization techniques have long supported comparisons of distributions and summary statistics. For example, v-plots combine local frequencies, global shapes, and aggregated statistics within familiar chart forms [BDL*20], and prior work on Likert-scale visualization clarifies design considerations for ordinal evaluation data [SSV*22]. Other approaches address conjunctions, ordering, and uncertainty. Prior work has shown how uncertainty can be embedded into parallel coordinates, for example through probabilistic extensions [BIL25], and has emphasized the role of visualization throughout Bayesian workflows [GSV*19]. Related work also includes methods for analyzing rankings and their changes over time [PKWQ20, WJN*23], as well as broader surveys of uncertainty visualization [Wei22].

While none of these works directly integrate Bayesian probabilities into conventional evaluation charts, they provide techniques and insights for comparing distributions, reasoning about uncertainty, and analyzing ordering. Building on this foundation, our work explores how Bayesian posterior probabilities can be incorporated into familiar evaluation charts.

5. Conclusion

We presented BPHC-based visualizations that augment NHST-based evaluation charts with Bayesian posterior probabilities while preserving their visual conventions. By expressing threshold-based confidence directly from posterior distributions, our approach conveys effect size, uncertainty, and confidence interpretably. We introduced scalar, one-dimensional, two-dimensional, and parallel-coordinates BPHC representations to support varying levels of detail and conjunction complexity. While scalar and parallel-coordinates BPHC report probabilities at a fixed threshold, the underlying framework directly supports evaluation across multiple thresholds, beyond the single-threshold examples shown here.

6. Acknowledgement

Microsoft 365 Copilot was used to improve language clarity and accelerate code creation for figures. This work was supported in part by JST CREST Grant Number JPMJCR20D4 and JST CRONOS Grant Number JPMJCS25K1, Japan.

References

[AGM19] AMRHEIN V., GREENLAND S., MCSHANE B.: Scientists rise up against statistical significance. *Nature 567* (2019), 305–307. 1

[BDL*20] BLUMENSCHEIN M., DEBBELER L. J., LAGES N. C., RENNER B., KEIM D. A., EL-ASSADY M.: v-plots: Designing hybrid charts for the comparative analysis of data distributions. *Computer Graphics Forum (EuroVis 2020) 39*, 3 (2020), 565–577. 4

[BIL25] BORRELLI G., ITTERMANN T., LINSEN L.: Mapping mental models of uncertainty to parallel coordinates by probabilistic brushing. *Computer Graphics Forum (EuroVis 2025) 44*, 3 (2025), 1–12. 4

[CG14] CORRELL M., GLEICHER M.: Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 2142–2151. 2

[GHNO02] GOTO M., HASHIGUCHI H., NISHIMURA T., OKA R.: RWC music database: Popular, classical, and jazz music databases. In *Proceedings of ISMIR 2002* (2002), pp. 287–288. 3

[GSV*19] GABRY J., SIMPSON D., VEHTARI A., BETANCOURT M., GELMAN A.: Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A 182*, 2 (2019), 389–402. 4

[HG14] HOFFMAN M. D., GELMAN A.: The No-U-Turn Sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research 15*, 1 (2014), 1593–1623. 1

[Kay24] KAY M.: ggdist: Visualizations of distributions and uncertainty in the grammar of graphics. *IEEE Trans. Vis. Comput. Graph. 30*, 1 (2024), 414–424. 2

[KNH16] KAY M., NELSON G. L., HEKLER E. B.: Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 4521–4532. 1

[KR12] KAPTEIN M., ROBERTSON J.: Rethinking statistical analysis methods for chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (2012), pp. 1105–1114. 1

[Kru10a] KRUSCHKE J. K.: Bayesian data analysis. *WIREs Cognitive Science 1*, 5 (2010), 658–676. 1

[Kru10b] KRUSCHKE J. K.: What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences 14*, 7 (2010), 293–300. 1

[Kru13] KRUSCHKE J. K.: Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General 142*, 2 (2013), 573–603. 2

[NG24] NAKANO T., GOTO M.: Using item response theory to aggregate music annotation results of multiple annotators. In *Proceedings of ISMIR 2024* (2024), pp. 1–9. 3

[O*23] OGATA K., ET AL.: The influence of Bouba- and Kiki-like shape on perceived taste of chocolate pieces. *Frontiers in Psychology 14* (2023), 1–13. 2, 3

[PKWQ20] PURI A., KU B. K., WANG Y., QU H.: RankBooster: Visual analysis of ranking predictions. In *EuroVis 2020 - Short Papers* (2020), pp. 1–5. 4

[Sam69] SAMEJIMA F.: Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* (1969). 3

[SSKQ25] SCHUBERT A.-L., STEINHILBER M., KANG H., QUINTANA D. S.: Improving statistical reporting in psychology. *Communications Psychology 3* (2025), 1–23. 2

[SSV*22] SOUTH L., SAFFO D., VITEK O., DUNNE C., BORKIN M. A.: Effective Use of Likert Scales in Visualization Evaluations: A Systematic Review. *Computer Graphics Forum (EuroVis 2022) 41*, 3 (2022), 43–55. 1, 4

[Toy24a] TOYODA H.: *Statistical Significance and the PHC Curve*, 1 ed. Springer Singapore, 2024. 1, 2

[Toy24b] TOYODA H.: *Statistics with posterior probability and a PHC curve*. Springer Singapore, 2024. 1, 2

[Wei22] WEISKOPF D.: Uncertainty visualization: Concepts, methods, and applications in biological data visualization. *Frontiers in Bioinformatics 2* (2022), 793819. 4

[WJN*23] WANG H., JIANG X., NAGARAJAN A., GUO X., DING L., WAN D., ZHAO J., CHEN Y.: Colorslope: a balanced visualization of overview and details on ranks over time. *Visual Intelligence 1*, 7 (2023), 1–13. 4

[WSL19] WASSERSTEIN R. L., SCHIRM A. L., LAZAR N. A.: Moving to a world beyond "$p < 0.05$". *The American Statistician 73*, sup1 (2019), 1–19. 1