Journal of Integrated Knowledge
Vol. 2 No. 7 (2026)

The Same Leader, Different Choice Sets: Ordinal Ties in Released Usability Feedback

Joseph Barker[1] · George Hughes[2] · Matthew Fisher[3]

[1] New Zealand · [2] Switzerland · [3] Netherlands

Published: July 08, 2026 | Article 863

Abstract

Three feedback instruments can name the same leader without identifying the same choice. In twelve complete four-product response blocks, SUS, an adjective rating and a recommendation item all kept RIT among Borda co-winners under every permitted resolution of observed ties. Only the adjective responses also allowed TRU to join it; TRU could not win alone. After individual participant omissions, seven of twelve adjective profiles had no necessary co-winner, while SUS and recommendation retained RIT throughout. Integer feasibility and independent dynamic programming agreed across all 52 analysis branches and cohorts. These exact choice sets describe a declared aggregation rule, without validating product quality or estimating population confidence. Keywords: usability feedback; ordinal ties; possible winners; Borda aggregation; identification bounds

The Same Leader, Different Choice Sets: Ordinal Ties in Released Usability Feedback

Abstract

Three feedback instruments can name the same leader without identifying the same choice. In twelve complete four-product response blocks, SUS, an adjective rating and a recommendation item all kept RIT among Borda co-winners under every permitted resolution of observed ties. Only the adjective responses also allowed TRU to join it; TRU could not win alone. After individual participant omissions, seven of twelve adjective profiles had no necessary co-winner, while SUS and recommendation retained RIT throughout. Integer feasibility and independent dynamic programming agreed across all 52 analysis branches and cohorts. These exact choice sets describe a declared aggregation rule, without validating product quality or estimating population confidence.

Keywords: usability feedback; ordinal ties; possible winners; Borda aggregation; identification bounds

1. From a reported experience to a collective choice

A product-feedback instrument can support an ordering without determining a unique collective choice. Several users may select the same response category for different products. An analyst who turns those categories into strict ranks must then decide what to do inside each tied block. Selecting one convenient order, averaging occupied ranks and preserving every compatible order are distinct operations. Their consequences cannot be inferred from the identity of a single displayed leader. This study investigates that distinction using complete repeated responses to four products, rather than comparing unpaired group averages or estimating a new usability factor.

The System Usability Scale was developed as a compact assessment of perceived usability (Brooke, 1996). Its creator's retrospective places its use within practical evaluation constraints and the distinction between subjective opinions and observed performance (Brooke, 2013). Lewis (2018) reviews its subsequent development as an instrument for perceived usability. These accounts motivate keeping the interpretation attached to the source responses. A higher ordinal position under a chosen rule does not establish that a product is objectively better, completes more tasks correctly or should replace a product intended for a different purpose.

The registered abstract of Yuan et al. (2026) connects experience measurement, scale validation, uncertainty-aware differences and whether evidence supports decisions. Our analysis addresses the last link: does a stated ordinal aggregation actually identify its decision when the source responses contain ties? That paper's nonpublic community data and full methods were not accessed. We implement neither its Bayesian pooling nor its privacy framework. Its relevance is the requirement to examine what an experience indicator warrants before using it for a decision; the present empirical evidence comes entirely from a different public numeric release.

Several conceptual boundaries matter. Stevens (1946) distinguishes ordering from stronger numerical scale properties, and Liddell and Kruschke (2018) examine consequences of applying metric models to ordinal observations. We avoid equating distances across the three instruments. Messick (1995) locates validity in justified interpretations and uses of assessment evidence; Jacobs and Wallach (2021) similarly make operationalization assumptions explicit. Here the operation is fixed Borda aggregation of within-person orders. Its identified outcome is a property of those recorded responses and that rule, rather than validation of the underlying experiences or a new measure of product utility.

2. A public repeated-feedback collection

Berke et al. (2019) published an institutional supplement with response values supporting research on usability-questionnaire translations. We use only its complete English CSV, acquired and hashed before analysis. The associated supplement and actual English answer sheet were read, together with the registered article abstract and metadata. The full ACM article page was inaccessible, so we do not claim to have reviewed its complete methods or to have discovered a relationship absent from that paper. The present question concerns joint choice identification under tie completions, distinct from general instrument association or translation validation.

The English file contains 58 product-response rows and 22 participant codes. Twelve codes each evaluate all four artifacts: the RIT website, Monroe County government website, Trulia Android application and LLION digital camera. Ten older codes contain only a RIT evaluation and no recommendation response. The supplement explains that these earlier participants were not asked that item. Their absence is structural to the study record. Missing-data reasoning requires retaining such a distinction rather than treating a blank as a neutral or unfavorable rating (Little and Rubin, 2019). All 58 rows remain in the source flow; the fixed collective-choice analysis uses the twelve complete four-product blocks, totaling 48 rows.

The source has no duplicated participant-artifact key. Every observed item falls within its documented response domain. All ten SUS items and both other usability fields are present on every source row, while ten recommendation values are absent. The actual CSV labels its participant field PID and its adjective field SUS-Adj, differing from shorthand in the webpage description. We used the actual headers after a visible initial parse failure, preserving source bytes. Dataset documentation should identify composition, collection and intended uses (Gebru et al., 2021); those details determine which repeated comparisons this release can support.

Three instruments supply separate ordinal signals. SUS combines ten five-category responses, with alternating item polarity. We independently reconstruct its score by summing the declared item contributions and multiplying by 2.5. The adjective item rates overall user-friendliness on seven ordered categories. The recommendation item asks likelihood of recommending the product to a friend or colleague on a zero-to-ten scale. Recommendation has a separate commercial history (Reichheld, 2003). Keiningham et al. (2007) examine recommendation metrics against firm revenue growth, a substantially different endpoint. We use the raw item response, without calculating an aggregate Net Promoter Score index or inferring an observed recommendation, purchase or revenue outcome.

The four artifact identifiers refer to the particular historical evaluations, not current versions of the websites or devices. The artifacts serve different functions, so this conditional choice exercise cannot justify replacing one with another for a shared practical task. Nor do we compare participant language groups or infer personal characteristics from the release. Public participant codes are enough to join repeated records; they are not an anonymization certificate. The institutional page invites inspection of the CSV, but a general redistribution license for the original data has not been established. The reproducibility record therefore supplies original-source links, hashes, code and derived aggregate bounds without claiming a new license over the raw CSV or questionnaire PDF.

3. What the fixed decision rule identifies

Social choice examines how individual orderings produce collective outcomes (Arrow, 2012), and computational work studies the algorithms needed to obtain them (Brandt et al., 2016). We select a single rule rather than searching for a rule that produces a preferred product. With four products, each strict order receives Borda points three, two, one and zero, respectively, as in the positional-scoring formulation described by Walsh (2007). Every participant contributes six points. Equal participant weighting defines this finite decision, without estimating selection probabilities. The broader question of what weighting is intended to accomplish remains important (Solon et al., 2015); our weights are an explicit rule choice, not a population correction.

Each instrument first defines a weak order within each participant. Strictly higher responses must stay above lower responses. Products receiving an equal response can appear in any order within that tied block. We enumerate every strict order satisfying those conditions. These completions are permitted encodings of the recorded information, not estimates of hidden genuine preferences. Kendall (1945) treats ties as a substantive ranking issue; assigning an arbitrary strict position would instead add information that these responses do not supply. Our question is conditional on permitting such completions, rather than asserting that a respondent really distinguishes every tied product.

The possible and necessary winner framework separates existence from guarantee (Konczak and Lang, 2005; Xia and Conitzer, 2011). A possible co-winner reaches the largest aggregate score in at least one permitted joint completion. A possible unique winner is strictly ahead in at least one. A necessary co-winner remains among the largest-score products in every completion. Necessary co-winners can coexist with another possible co-winner; they need not be uniquely first throughout. Proposition P1 states the inclusion of every necessary co-winner in the possible co-winner set. We check that inclusion in every analysis rather than confusing the three definitions in a single winner label.

We also calculate a tie-average baseline. Each tied product receives the average of its occupied Borda points, and those points are summed across participants. This convention yields one score vector, which can be useful as a display. It does not characterize every compatible strict completion and need not itself be one such integer vector. Rank-aggregation research offers other ways to combine rankings (Dwork et al., 2001). None is evaluated here. Keeping the rule fixed isolates the uncertainty about order inside observed categories from uncertainty introduced by switching aggregation methods.

For every product we retain its minimum and maximum attainable total, and for every product pair its aggregate margin bounds. Marginal bounds for two products need not be jointly attainable. Winner feasibility is therefore tested against the complete joint constraints rather than inferred by combining favorable endpoints from separate intervals. Identified sets distinguish conclusions warranted by available information from points obtained under extra assumptions (Manski, 2003). The bounds here are exact for this finite response collection and completion definition. They have no repeated-sampling coverage interpretation of the kind considered by Neyman (1937), and we attach no confidence percentage to them.

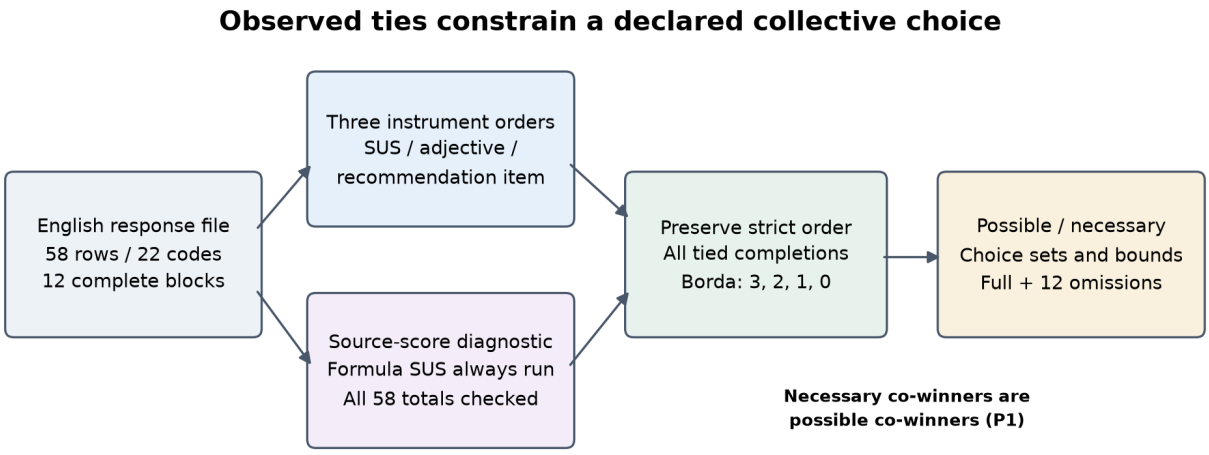

Figure 1. Original identification model. Three instruments plus the always-executed formula-SUS diagnostic enter the same fixed completion and aggregation procedure. P1 concerns logical set inclusion. Arrows indicate calculation, without a causal or latent-preference claim.

4. Exact computation and complete sensitivity analysis

The local protocol froze the population, three primary instruments, diagnostic formula branch, point schedule and all deletion calculations before reconstructing scores or computing ranks. The primary program enumerates the 24 possible product permutations and retains those respecting each participant's inequalities. Possible co-winner and unique-winner queries then become finite integer-feasibility problems with one permitted score vector per participant. Depth-first search uses exact integer margin bounds and memoization. Every feasible answer has a saved permitted witness whose aggregate inequalities are checked; infeasible answers remain explicit.

A separate program reparses the original CSV, builds ordered tied blocks directly and permutes labels within those blocks. Dynamic programming enumerates every reachable aggregate score vector, using a reversed artifact and participant namespace. Because the four totals always sum to six times the participant count, three coordinates determine the fourth. All winner classifications, marginal score bounds, pair margins and baseline scores are recomputed from these reachable vectors. This implementation does not import primary analysis functions or reuse its permitted-order lists. The agreement therefore checks two constructions of the same declared finite problem.

Each primary instrument and the formula-SUS diagnostic is evaluated on the complete twelve-person cohort and all twelve cohorts omitting exactly one participant. Thus 52 branch-cohort combinations remain, with no favorable deletion selected for reporting. Deletions leave eleven complete participants and a total of 66 Borda points, compared with 72 in the complete cohort. These are deterministic changes to composition. They do not model new users, recover a sampling distribution or simulate an independent repetition of the original study. An empty necessary-winner set and unchanged classifications are equally reportable outcomes.

The source-score diagnostic is always executed. All 58 published SUS values match their exact formula reconstructions, and the reconstructed branch reproduces every full and deletion result. No correction is needed. The independent programs agree on all 208 product records and 312 pair-margin records, including 2,080 integer bound and flag comparisons. All 119 feasible witnesses pass independent checks against the original permitted orders and their claimed inequalities. There is no algorithm disagreement or

numerical discrepancy. These are computational verification results, distinct from external validation of the questionnaire or its interpretations.

5. Shared leadership with different admissible choices

The three instruments provide different amounts of ordinal freedom (Table 1). SUS has only one participant with a tied product pair, yielding two joint completions. The adjective and recommendation items each have five participants with ties, but yield 384 and 192 joint completions, respectively. Multiple completions can produce the same aggregate scores: the independent dynamic program finds 108 distinct adjective vectors and 44 recommendation vectors. These are combinatorial counts, without an assumption that the completions are equally likely or that a larger count implies a worse instrument.

Table 1. Complete-cohort information support. Each instrument has twelve participants and 72 within-person product pairs. Joint completions choose one compatible strict order per participant; aggregate vectors count distinct total-score outcomes.

Instrument	Participants with ties	Tied product pairs	Joint completions	Aggregate vectors
SUS	1 / 12	1 / 72	2	2
Adjective	5 / 12	10 / 72	384	108
Recommendation item	5 / 12	8 / 72	192	44

RIT remains a necessary co-winner under every instrument in the complete cohort. Its SUS total is fixed at 29 Borda points, compared with TRU's fixed 24; the only SUS freedom changes scores of lower products. Recommendation allows a wider RIT range, from 29 to 33, but still cannot remove it from the winning set. This is an important null result for an account that expects any coarse response scale to destroy the collective decision. More permissible orders do not automatically create a different possible winner.

Table 2. Exact full-cohort aggregate score extrema and tie-average baseline, N=12 participants. All scores are Borda points under the fixed 3-2-1-0 schedule. Each row is a marginal range; endpoints in different rows are not asserted jointly attainable.

Instrument	Artifact	Attainable score extrema	Tie-average baseline
SUS	CAM	[14, 15]	14.5
SUS	MC	[4, 5]	4.5
SUS	RIT	[29, 29]	29
SUS	TRU	[24, 24]	24
Adjective	CAM	[12, 17]	14.5
Adjective	MC	[5, 8]	6.5
Adjective	RIT	[25, 32]	28.5
Adjective	TRU	[20, 25]	22.5
Recommendation item	CAM	[11, 16]	13.5
Recommendation item	MC	[5, 6]	5.5
Recommendation item	RIT	[29, 33]	31
Recommendation item	TRU	[19, 25]	22

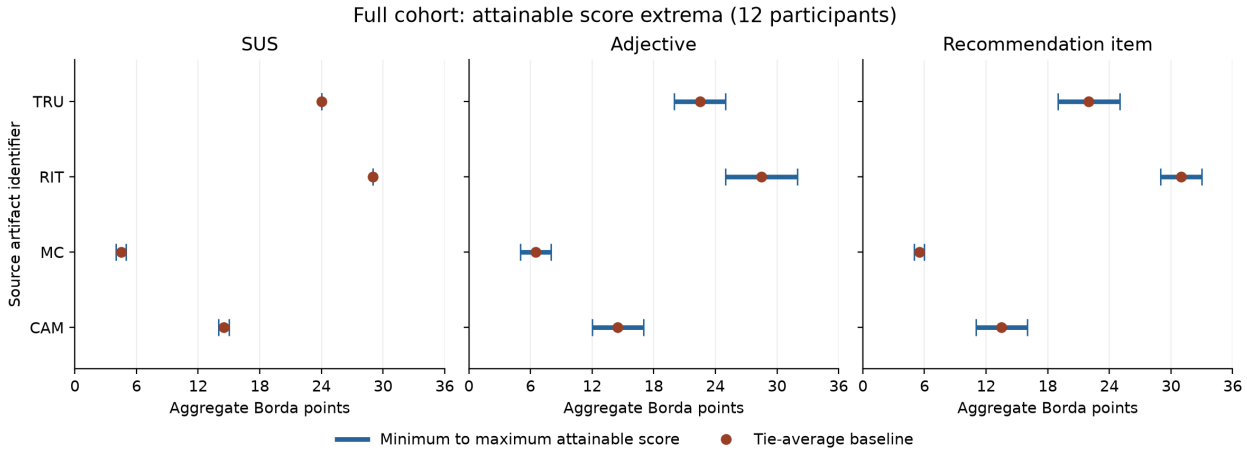

Figure 2. Full-cohort minimum and maximum scores on a common zero-to-36 axis. Dots show the stated baseline, and segments connect extrema without asserting a continuous distribution or simultaneous attainability across products. These are identification bounds, not confidence intervals.

The adjective instrument supplies the qualified case. Its RIT-TRU aggregate margin ranges from zero to twelve points. A compatible completion can therefore make them co-winners, although none makes TRU a unique winner. RIT belongs to every winning set, while TRU belongs to only some. The baseline scores, 28.5 for RIT and 22.5 for TRU, show a six-point difference yet conceal that boundary tie. By comparison, the corresponding margin is exactly five under SUS and ranges from four to fourteen under the recommendation item. The full result concerns the location of admissible uncertainty relative to the winning boundary, not the width of score ranges alone.

Table 3. Full-cohort choice sets and complete deletion sensitivity. The final column counts omission cohorts with no necessary co-winner; its denominator is the twelve distinct single-participant omissions. Possible and necessary sets refer to all compatible joint completions.

Instrument	Possible co-winners	Possible unique winners	Necessary co-winners	Empty necessary set after omission
SUS	RIT	RIT	RIT	0 / 12
Adjective	RIT, TRU	RIT	RIT	7 / 12
Recommendation item	RIT	RIT	RIT	0 / 12

Participant deletion changes only the adjective choice classification. Seven of its twelve omission cohorts have no necessary co-winner, and both RIT and TRU can then win uniquely. In the remaining five, RIT is the sole possible and necessary co-winner. The RIT-TRU margin lower bound is minus one in each of the seven ambiguous cohorts, while their upper bounds remain positive. Under SUS, all twelve omissions retain only RIT, with its margin over TRU ranging from three to six points across omissions. Recommendation also retains only RIT throughout; its omission-specific lower margin bounds range from two to six. No deletion gives CAM or MC a possible co-winning position.

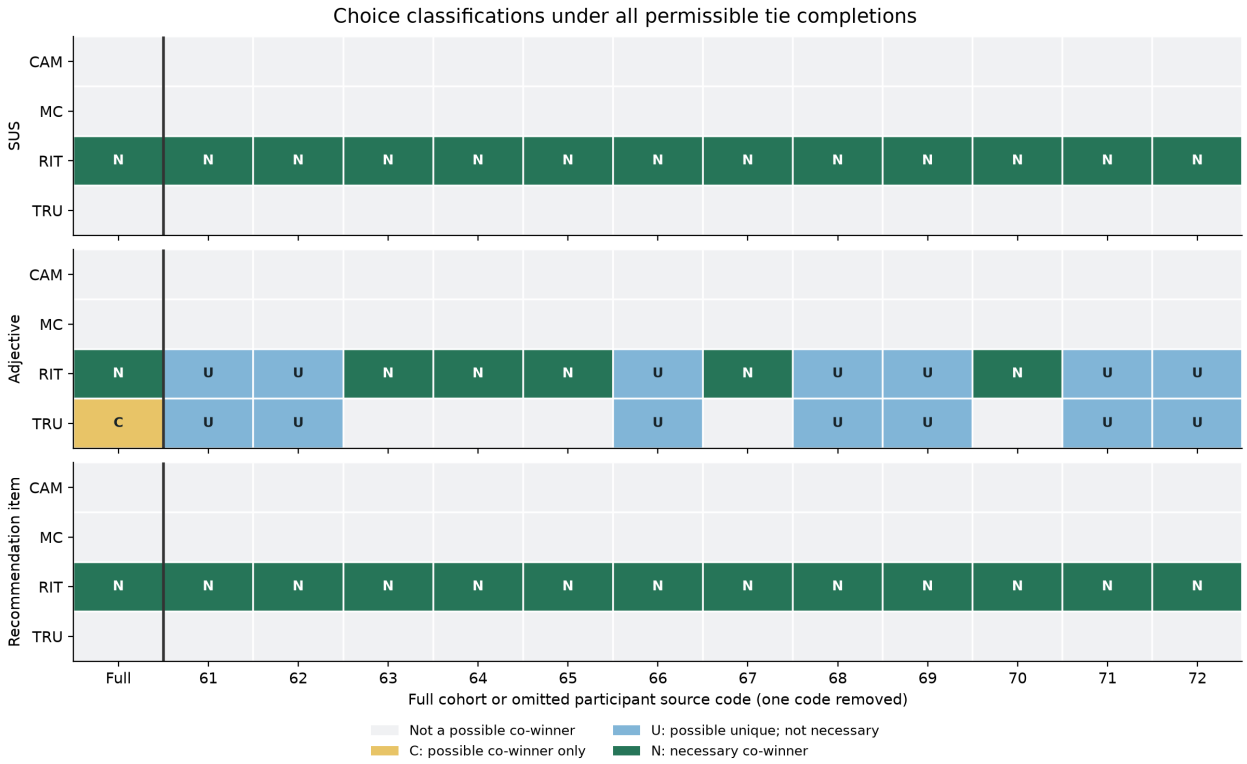

Figure 3. All full and deletion classifications for the three primary instruments. N marks a necessary co-winner; U a possible unique winner without necessity; C a possible co-winner without unique-winning feasibility. Gray cells retain products that cannot co-win. Formula-SUS results match the SUS panel in every cohort.

6. Interpreting a decision set without overstating it

The empirical distinction is between a shared baseline leader and the stronger claim that a particular collective choice is identified. All three complete profiles share RIT, but the adjective profile admits an additional co-winner. Its deletion analysis further shows how a modest composition change can remove every guaranteed co-winner. Neither finding establishes that the adjective instrument mismeasures usability. It provides fewer categories and asks a different question from the ten-item composite; preserving its ties is part of respecting the information actually recorded.

A tie-average display is also not erroneous merely because a completion set is larger. It is a fully specified convention. The interpretive error would arise from presenting its one output as a guarantee over other permitted within-category orders. Conversely, the possible presence of TRU in a set does not assign it a probability, document an unobserved preference or establish that a product manager should select it. The rule was chosen to make one ordinal decision inspectable. An application with other consequences would need to justify that rule, its participant weights, the candidate set and its evidence requirements independently.

Data fitness includes what a representation permits a consumer to conclude, beyond whether stored numbers are accurate (Wang and Strong, 1996). Here all source SUS arithmetic reconciles, yet the choice question still differs across instruments and composition states. A numerical consistency check therefore cannot replace interpretation of the item, the unit receiving weight or the information added by a tie convention. Our results favor reporting the possible and necessary sets beside a single baseline when that distinction matters for the declared decision.

Reproducibility records bind each source object, eligibility decision and derived bound. Buneman et al. (2001) distinguish tracing data origins and transformations, while Herschel et al. (2017) discuss provenance for reproducibility and accountability. The saved aggregate tables, source hashes, protocol and integer witnesses support this tracing. NumPy organizes the figure matrices (Harris et al., 2020), and Matplotlib generates the original diagram and numerical displays (Hunter, 2007). Scientific Agent Skills procedures supported writing, statistical interpretation and figure review (Kassis et al., 2026). These tools assist inspectable work; they supply no certification of instrument validity.

Several limits are decisive. The complete cohort contains only twelve participants from one historical release, evaluating unlike artifacts in the source study's conditions. There is no probability sample of users or products, no rerun and no behavioral outcome. Permitted strict completions are an analysis of information under a specified convention, not a latent preference model. Public numeric access does not establish raw-data relicensing or anonymity. Executable agreement supports computational reproducibility in the sense discussed by Peng (2011), while leaving broader replication and accountable scientific judgment unresolved. The result is a narrow one: the same observed leader can coexist with different exact choice sets, and some of those sets depend on one person's inclusion.

Data, code and review status

Original English responses and the questionnaire remain available from the authors' institutional supplement. Companion artifacts provide derived aggregate tables, both exact algorithms, the frozen protocol, source hashes and evidence-review records. Original CSV/PDF redistribution is not represented as newly licensed. Human authorship, declarations, scientific approval and source verification remain **PENDING**; submission readiness is **false**. No private community data, new respondents, model calls or product executions were obtained.

References

1. Harris, Charles R.; Millman, K. Jarrod; van der Walt, Stéfan J.; Gommers, Ralf; Virtanen, Pauli; Cournapeau, David; Wieser, Eric; Taylor, Julian; Berg, Sebastian; Smith, Nathaniel J.; Kern, Robert; Picus, Matti; Hoyer, Stephan; van Kerkwijk, Marten H.; Brett, Matthew; Haldane, Allan; del Río, Jaime Fernández; Wiebe, Mark; Peterson, Pearu; Gérard-Marchant, Pierre; Sheppard, Kevin; Reddy, Tyler; Weckesser, Warren; Abbasi, Hameer; Gohlke, Christoph; Oliphant, Travis E.. (2020). Array programming with NumPy. Nature. https://doi.org/10.1038/s41586-020-2649-2
2. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677
3. Keiningham, Timothy L.; Cooil, Bruce; Andreassen, Tor Wallin; Aksoy, Lerzan. (2007). A Longitudinal Examination of Net Promoter and Firm Revenue Growth. Journal of Marketing. https://doi.org/10.1509/jmkg.71.3.039
4. Hunter, John D.. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. https://doi.org/10.1109/mcse.2007.55
5. Brooke, John. (2013). SUS: A Retrospective. Journal of Usability Studies. https://www.researchgate.net/publication/285811057_SUS_a_retrospective
6. Walsh, Toby. (2007). Uncertainty in Preference Elicitation and Aggregation. Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence. https://aaaipress.org/Library/AAAI/2007/aaai07-001.php
7. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847
8. Liddell, Torrin M.; Kruschke, John K.. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. Journal of Experimental Social Psychology. https://doi.org/10.1016/j.jesp.2018.08.009
9. Brandt, Felix; Conitzer, Vincent; Endriss, Ulle; Lang, Jérôme; Procaccia, Ariel D. (eds.). (2016). Handbook of Computational Social Choice. Cambridge University Press. https://doi.org/10.1017/CBO9781107446984
10. Herschel, Melanie; Diestelkämper, Ralf; Ben Lahmar, Houssem. (2017). A survey on provenance: What for? What form? What from?. The VLDB Journal. https://doi.org/10.1007/s00778-017-0486-1
11. Buneman, Peter; Khanna, Sanjeev; Tan, Wang-Chiew. (2001). Why and Where: A Characterization of Data Provenance. Database Theory - ICDT 2001, 316-330. https://doi.org/10.1007/3-540-44503-X_20.
12. Berke, Larwan; Huenerfauth, Matt; Patel, Kasmira. (2019). Design and Psychometric Evaluation of American Sign Language Translations of Usability Questionnaires. ACM Transactions on Accessible Computing. https://doi.org/10.1145/3314205
13. Kendall, M. G.. (1945). THE TREATMENT OF TIES IN RANKING PROBLEMS. Biometrika. https://doi.org/10.1093/biomet/33.3.239
14. Brooke, John. (1996). SUS: A Quick and Dirty Usability Scale. Usability Evaluation in Industry. https://doi.org/10.1201/9781498710411-35
15. Lewis, James R.. (2018). The System Usability Scale: Past, Present, and Future. International Journal of Human-Computer Interaction. https://doi.org/10.1080/10447318.2018.1455307
16. Arrow, Kenneth J.. (2012). Social Choice and Individual Values. Yale University Press. https://yalebooks.yale.edu/book/9780300186987/social-choice-and-individual-values/
17. Little, Roderick; Rubin, Donald. (2019). Statistical Analysis with Missing Data, Third Edition. Wiley Series in Probability and Statistics. https://doi.org/10.1002/9781119482260
18. Reichheld, Frederick F.. (2003). The One Number You Need to Grow. Harvard Business Review. https://hbr.org/2003/12/the-one-number-you-need-to-grow
19. Manski, Charles F.. (2003). Partial Identification of Probability Distributions. Springer New York. https://doi.org/10.1007/b97478
20. Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; III, Hal Daumé; Crawford, Kate. (2021). Datasheets for datasets. Communications of the ACM. https://doi.org/10.1145/3458723
21. Kassis, Timothy; Agarwal, Vinayak; He, Yuhuan; Patel, Darshil; Brueckner, Aubrey M.. (2026). Scientific Agent Skills: A Library of Procedural Knowledge for Research Agents. arXiv preprint. https://arxiv.org/abs/2609.00065
22. Dwork, Cynthia; Kumar, Ravi; Naor, Moni; Sivakumar, D.. (2001). Rank aggregation methods for the Web. Proceedings of the 10th international conference on World Wide Web. https://doi.org/10.1145/371920.372165
23. Konczak, Kathrin; Lang, Jérôme. (2005). Voting Procedures with Incomplete Preferences. IJCAI2005Workshop on Advances in Preference Handling. https://www.lamsade.dauphine.fr/~lang/papers/konlan05a.pdf

24. Jacobs, Abigail Z.; Wallach, Hanna. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445901

25. Yuan, Yuewei; Huang, Jialei; Yin, Jiayang; Xu, Tianyi. (2026). Visual Verification of user Experience Differences under Privacy Constraints in Anonymous Feedback Data. SSRN working paper 7393298. https://doi.org/10.2139/ssrn.7393298.

26. Solon, Gary; Haider, Steven J.; Wooldridge, Jeffrey M.. (2015). What Are We Weighting For?. Journal of Human Resources. https://doi.org/10.3368/jhr.50.2.301.

27. Neyman, J.. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences. https://doi.org/10.1098/rsta.1937.0005

28. Messick, Samuel. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.. American Psychologist. https://doi.org/10.1037/0003-066x.50.9.741

29. Wang, Richard Y.; Strong, Diane M.. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. https://doi.org/10.1080/07421222.1996.11518099

30. Xia, Lirong; Conitzer, Vincent. (2011). Determining Possible and Necessary Winners under Common Voting Rules Given Partial Orders. Journal of Artificial Intelligence Research. https://doi.org/10.1613/jair.3186