Similar Ratings, Different Consensus: Visualizing Agreement Uncertainty in Survey Data

Vrushali Koli, Yi Meng, Eugene P. Deess and Aritra Dasgupta *

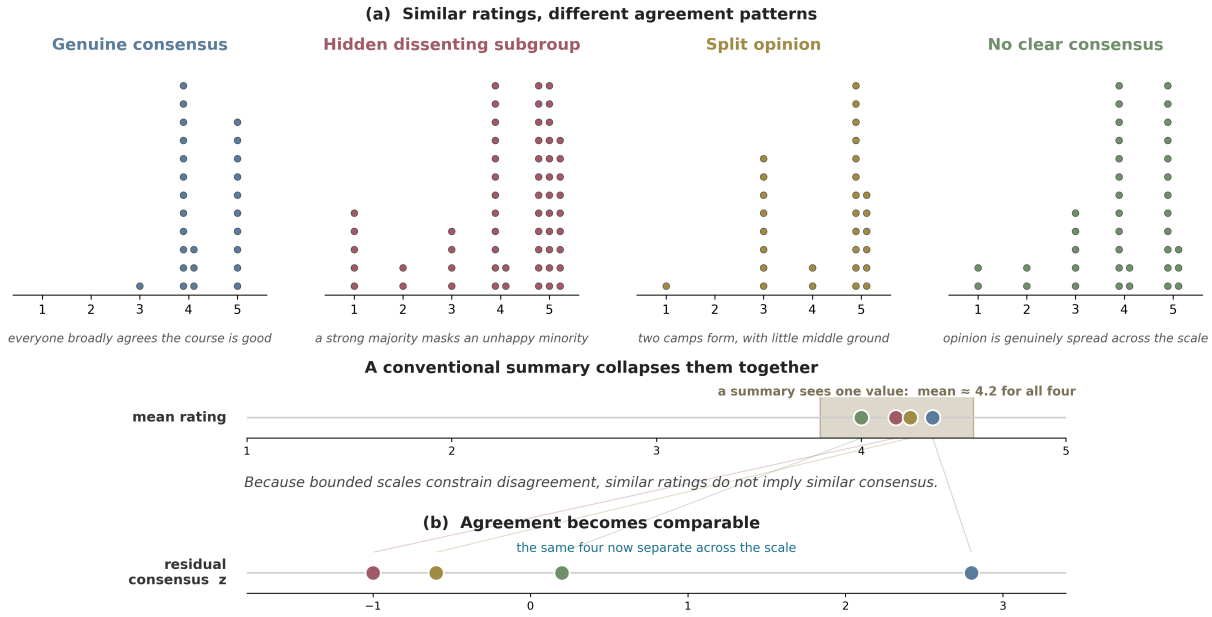


Figure 1: **Summary statistics hide how much a group agrees.** Four real courses at nearly the same average rating (about 4.0 to 4.4 on a five-point scale) show very different agreement: genuine consensus, a hidden dissenting subgroup, split opinion, and no clear consensus (a). The figure previews the argument in one view: a conventional summary places the four together, since they cluster by mean rating, while residual consensus separates them, since they spread by z (b). Color links each course from its distribution to its position on the two axes.

Abstract

Survey results are routinely summarized with scalar measures of agreement to compare questionnaire items, yet those summaries are hard to interpret on bounded rating scales. Most uncertainty visualizations communicate uncertainty by representing multiple possible values, estimates, or outcomes. We consider a complementary source: the responses and the computed agreement score are fixed, but the meaning of that score depends on where it sits on the scale. Because responses have less room to spread near the endpoints of a five-point Likert item, agreement there is mechanically constrained, so the same score can arise from very different distributions and can imply different consensus at different rating levels. We call this ambiguity *agreement uncertainty*; collecting more responses does not resolve it, since the score must instead be read against the feasible agreement space at its rating level. We introduce *residual consensus*, a calibration that expresses agreement relative to what the scale allows, informative only where that space has enough resolution and therefore read alongside the response distribution itself. Drawing on ideas from possibility-framing and meta-uncertainty in uncertainty visualization, we contribute a coordinated view that shows calibrated agreement, its positional informativeness, and the distribution together, demonstrated on real-world survey datasets to support more meaningful comparisons across bounded scales.

Index Terms: Uncertainty visualization, survey data, ordinal data, consensus, agreement, visual comparison, coordinated views.

1 Introduction

Surveys are read by assessing agreement: for each rated item, how united or divided were the respondents? For a single Likert item, the decision-relevant question is often not where the average sits but how much respondents agreed, and yet the summaries that report agreement can hide it. Figure 1 shows four real course-evaluation items with similar average ratings but very different agreement. Every conventional summary places them close together, yet what they imply about consensus, and the action they warrant, differs sharply. The small differences in these averages are not what makes their agreement diverge. Our data are institutional course evaluations, 60,700 item-level response distributions on a five-point scale; within them, more than 5,000 pairs of items have exactly the same mean, standard deviation, median, and interquartile range yet differ sharply in their response distributions, the survey counterpart of datasets with matched summary statistics but different shapes [1, 22]. Nor is the coupling incidental: ratings pile toward the top of the scale, where agreement is mechanically forced, so among the high-rated items that reviews focus on, raw agreement correlates with the rating level at about 0.97 and, read on its own, largely restates the mean; for 42% of items the scale leaves too little room to disagree for agreement to be read at all.

*All authors are with New Jersey Institute of Technology. E-mail: {vk435 | yi.meng | eugene.p.deess | aritra.dasgupta}@njit.edu.

Table 1: The four datasets and their roles. The two psychometric sets are probes of the metric's behavior across the scale, not target domains for interpretation.

Dataset	k	Scale position	Role
Course evals, primary	5	ceiling-heavy	motivating case
Course evals, 2nd institution	5	interior	robustness of culture, instrument
Big Five inventory	5	interior	probe across the scale
Rosenberg Self-Esteem	4	interior	probe across the scale

The difficulty runs deeper than one shared statistic, and it is specific to *bounded scales*: rating scales with a fixed set of ordered categories and hard endpoints, such as a 1-to-5 Likert item, on which a response can never fall below the lowest category or above the highest. A scale can express only so much agreement, and how much is not constant across it. Near an endpoint, responses cannot spread, so an item rated near the top is mechanically pushed toward the appearance of agreement. An item at the center can span the full range.

There are two implications of this observation. Identical agreement scores can arise from very different distributions, and, more subtly, the same score means different things at different rating levels. Scores are therefore not comparable across the scale: a high score near the ceiling is not the same claim as that score at the center.

What is uncertain here is not the data, which are fully observed, but how the score should be read. *Agreement uncertainty* is uncertainty about how a fixed agreement score should be interpreted once fully observed responses have been summarized on a bounded scale. The four courses of Figure 1 show the ambiguity: at nearly the same rating, one reflects genuine consensus and another hides a dissenting subgroup, yet the score alone cannot tell them apart. Several interpretations remain defensible until the score is placed against the *feasible agreement space*, the range of agreement the scale can produce at that rating level. Collecting more responses may stabilize the distribution, but it does not remove this structural ambiguity.

The missing information is representational: the score is correctly computed, but incomplete without the feasible space that determines what it can mean. Where uncertainty visualization commonly represents multiple possible values, estimates, analyses, or outcomes, here the observations and the summary are fixed and only their interpretation varies, at the transformation stage where uncertainty is known to enter [24]. The question is how a visualization should communicate when a summary statistic changes meaning across the domain in which it is measured.

Although our work uses institutional course-evaluation surveys as its example, these observations generalize to Likert-scale surveys broadly. Our object of study is the bounded ordinal scale itself, common to product ratings, clinical and well-being instruments, and psychological questionnaires.

We contribute (i) a characterization of agreement uncertainty, the observation that agreement scores on bounded ordinal scales are not directly comparable because the feasible agreement space changes with rating level; (ii) residual consensus, a calibration that restores comparability by expressing agreement relative to that space, with a region of positional informativeness shown to generalize across four datasets; and (iii) a coordinated visualization derived from three properties of the residual, which shows that space, the positional informativeness of the calibrated score, and the response distribution together, so an analyst can read and compare agreement in context.

2 Related Work

We build on three strands of prior work: uncertainty visualization, survey and ordinal data visualization, and the measurement of agreement, none of which expresses agreement relative to what a bounded scale allows or makes that context visible.

2.1 Uncertainty visualization

Uncertainty visualization is a mature subfield with dedicated surveys and typologies [3, 13, 24]. A recurring finding is that summary encodings such as error bars can mislead, and that showing the distribution or the space of outcomes is safer [6, 12]. Possibility- and frequency-framing displays, quantile dotplots, and hypothetical outcome plots, make uncertainty concrete as discrete outcomes and improve lay and expert judgments alike [9, 14, 18, 19]. The encoding also governs how strongly uncertain values draw attention [7, 11], a choice we make when a calibrated score is near-degenerate. These accounts also chart where uncertainty enters an analysis: at acquisition, during transformation, and in the visualization itself, a representation with uncertainty of its own [8, 21, 26], and across the data, model, and interpretive stages a data worker moves through in practice [4]. Agreement uncertainty enters at the transformation stage, when fully observed responses are summarized into a score. Its nearest relative is the *disagreement* uncertainty of Skeels et al., where different conclusions are drawn from the same data [29]; we make one kind of disagreement comparable on a bounded scale, but ours arises from a single, correctly computed summary whose meaning shifts with scale position, not from competing conclusions or analyses [16, 17]. It is likewise neither the direct quantitative uncertainty of facts and numbers nor the indirect qualitative uncertainty of a model's fidelity, conveyed through subjective judgments such as forecaster confidence [25]; and it is neither the internal uncertainty of a reasoner's ignorance nor the external uncertainty of the world's dispositions [15], since the responses are fully observed and the ambiguity belongs to the representation. That makes it computable rather than subjective, since the feasible agreement space can be derived and the score calibrated against it, so what varies is a fixed summary's interpretation as that space changes across the scale. The open problem, though, is not to classify this uncertainty but to make it visible in a comparison task, which turns on how survey responses are displayed and how agreement is measured, the subjects of the next two subsections.

2.2 Visualizing survey and ordinal data

Survey responses are most often shown as bar or stacked-bar charts, and the diverging stacked bar, centered on the neutral category, is a standard for Likert items [10, 28]. A systematic review finds that such charts dominate visualization evaluations and that agreement structure is rarely encoded directly [30]. These displays faithfully communicate a single item's response distribution, as our design preserves, but they do not reveal how agreement changes with rating level, so agreement cannot be compared across items with different ratings. We keep the distribution in reach [24, 30] and add the calibrated position that makes cross-item comparison valid.

2.3 Measuring agreement and consensus

Ordinal agreement has a long literature on measurement. Dispersion- and entropy-based indices summarize how spread or concentrated responses are [2, 20, 31], agreement coefficients target ranked or rated choices [27], and ordinal agreement indices calibrate against maximal disagreement [32]; inter-rater reliability measures such as Fleiss' kappa quantify rater consistency rather than the shape of one item's responses. What unites the single-item measures is that on a bounded scale, each inherits the coupling to location we describe: a value that signals strong agreement near the ceiling is unremarkable at the center. None expresses agreement as a position within the feasible agreement space conditioned on rating level, which is the

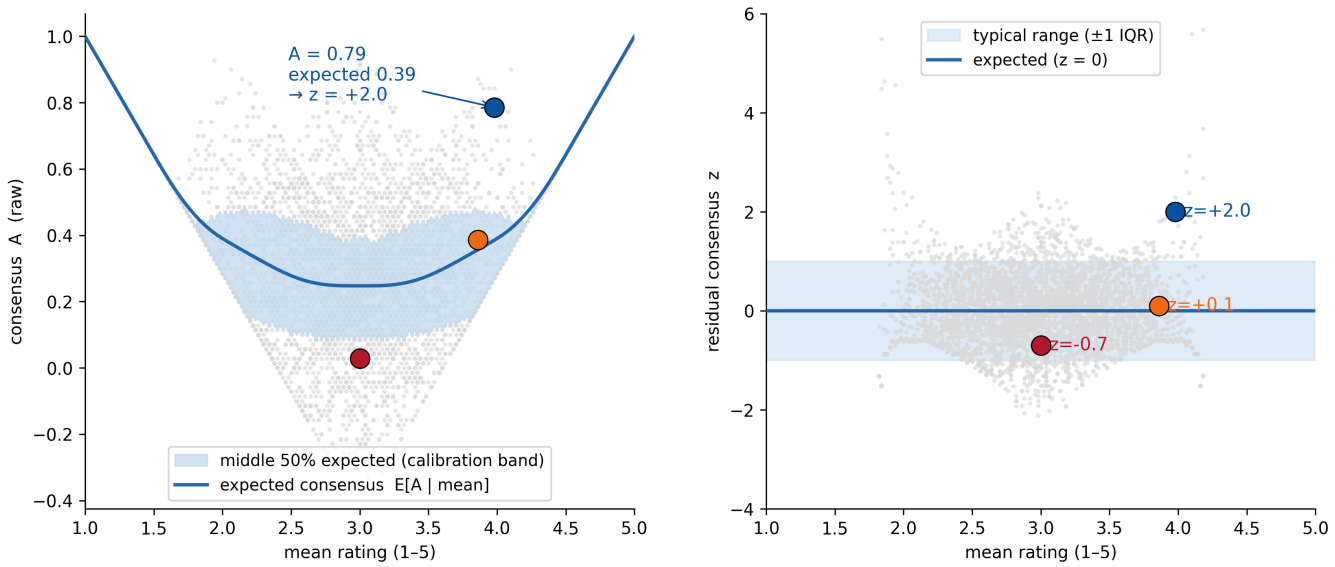


Figure 2: **Residual consensus makes agreement comparable across a bounded scale.** Left: the agreement a bounded scale can express is not constant. The feasible space pinches toward the ceilings, where responses have little room to differ, so a raw score means different things at different rating levels, and the three located items sit in bands of very different widths. Right: residual consensus reads each item by its position within the band the scale actually permitted, flattening the space so that a given value carries the same meaning at every rating level and items become directly comparable. This calibration is the argument's core: it turns an agreement score trapped by its rating level into one that reads consistently across the whole scale.

calibration we contribute, and none pairs such a calibration with a visualization that makes the resulting score legible and comparable. These gaps pose an analytical question, which the next section answers: how that space changes across the scale, and what properties any calibration of it must satisfy.

3 Characterizing Agreement Uncertainty

In this section, we characterize why agreement is not comparable across a bounded scale, a consequence of the scale's geometry rather than of measurement error, and derive the properties a faithful display must satisfy. We proceed in four steps: a base measure of agreement, the feasible-space geometry that makes it non-comparable, a calibration that restores comparability, and three empirical properties of the calibrated score that constrain the visual encoding of Section 4.

A base measure of agreement. Analyzing agreement needs a scalar that says how tightly a set of responses clusters, and three requirements narrow the choice. The measure should respect the order of the scale, since a rating is ordinal and not merely categorical; it should anchor on the median, the location statistic that suits a bounded ordinal scale and resists a lone distant response; and it should be bounded and directly interpretable. Rather than invent one, we borrow the coefficient of agreement of Riffenburgh and Johnstone [27], which meets all three. For counts $c_1, \ldots, c_k$ on a k-point scale with $n = \sum_j c_j$ responses and sample median m, write the mean absolute deviation from the median as $D = \frac{1}{n}\sum_j c_j\,|x_j - m|$. The coefficient is

$$A \;=\; 1 - \frac{D}{D_{\max}}, \tag{1}$$

where $D_{\max}$ is the largest deviation the scale can produce at that median. So $A = 1$ when every response sits at the median, and A falls toward 0 as mass moves to the extremes. Because A in (1) is a simple rescaling of the deviation D (high agreement where deviation is low), nothing that follows depends on the particular normalization; what matters is that agreement is measured from the median.

The alternatives make clear why this base is the right one. Variance and the standard deviation anchor on the mean, which is not the natural center of a bounded ordinal scale and which a single extreme rating can pull. Entropy-based consensus measures [31] capture dispersion but discard order, so they cannot separate a split between neighboring categories from a split between the two ends. Dedicated ordinal indices, such as van der Eijk's agreement [32] and Leik's consensus [20], do respect order, but none is framed against the feasible space that, as we show next, a bounded scale forces on any agreement score. A median-anchored coefficient is bounded, order-aware, and coherent with the median we already read as the item's location, which makes it the natural quantity to calibrate.

The feasible agreement space. This base measure is not comparable across the scale, and Figure 1 shows why: on a bounded scale, the meaning of agreement shifts with location. Enumerating every possible response distribution and plotting the achievable agreement against the mean yields a pinched envelope (Figure 2, left): near either endpoint the achievable range collapses toward its ceiling, because a near-extreme mean leaves almost no room to disagree. This space is wide at the center and narrow at the ends, so an agreement value that is unremarkable at the center is nearly forced at the ceiling.

This geometry has a direct empirical signature. On real evaluation data, where ratings cluster high, raw agreement correlates with the mean at about 0.97; read alone, it largely restates the rating level. The gallery of Figure 1 is this pinch in practice: its genuine-consensus and dissenting-subgroup courses sit at nearly the same rating yet on opposite sides of what the scale allows, so their agreement scores are not two points on one axis but two different claims.

Calibrating against the feasible space. We restore comparability by calibrating each item against this feasible space. *Residual consensus* is the observed agreement A minus the expected agreement at the item's rating level and sample size, divided by the feasible spread there. The reference is stated plainly: conditional on the mean and respondent count, every response distribution consistent

with that mean is treated as equally likely, a uniform reference over feasible outcomes rather than a fitted model. The residual is the item's position within the band of agreement the scale permits, and it supports calibrated comparison wherever the band remains wide enough to resolve it. After calibration, raw agreement correlates with the mean at 0.97, while the residual correlates at about 0.2, so items at different rating levels become comparable (Figure 2, right). Comparability here is concrete: where the feasible band has width, two items with the same residual show the same level of agreement regardless of rating level. The residual is invariant to sample size under fixed proportions, which is why it describes distributional structure rather than sampling error, and why relating a score to feasible agreement is a matter of interpretation rather than a statistical estimate.

Three properties that constrain any encoding. A calibrated score restores comparability, but it is legible only if the display respects the following three properties of the residual, each an empirical finding that constrains the visual encoding that follows.

R1: It is a qualifier, not a verdict. A residual of -1 means nothing until one knows whether the center sits at excellent or at poor; the same low agreement is good news at a high median and bad at a low one (Figure 3). The median is the right companion metric because it is robust to a dissenting minority and coheres with the median-anchored measure. Two mirror-image leans can share the same mean, standard deviation, and agreement score, so consensus cannot separate them, yet their medians differ and the median can. In the opposite case, where the median is uninformative, the residual is what discriminates. The display must pair the two.

R2: It adds information beyond variance only in part of the scale. On the screened corpus (items with a mean rating of 4.0 or above, the apparently high-rated cluster where mean ratings can hide lack of consensus) the residual and mean-adjusted standard deviation correlate at only -0.37 and flag disjoint worst-ranked sets: the residual flags broadly positive but internally divided cohorts (the split and dissenting-subgroup courses of Figure 1), variance flags hard dissenting blocs, and neither finds the other's. Entropy does not substitute either, and for a systematic reason: it, too, changes meaning across the scale, correlating with the mean at about -0.97.

But -0.37 is a pooled value, and the split by rating level shows what it hides. In the interior, the residual is nearly redundant with variance (Spearman -0.90); at the ceiling, it degenerates toward zero and carries no usable signal ($+0.01$); only at the shoulder between them, where variance loses resolution while the scale still permits disagreement, does it retain defined and independent signal (-0.72). The residual's added value is therefore regional, and the display must show where that value lies rather than assume that it holds everywhere.

R3: Its informativeness is positional. Whether the residual is even defined, and how finely it discriminates, depends on where a survey sits in the scale. Our institutional course evaluations are the case study, the survey data whose interpretation motivates this work. The two psychometric surveys, the Big Five inventory across its five factors and the Rosenberg Self-Esteem Scale (Table 1), are not domains of interest but probes. We include them to test whether the metric behaves across the whole scale, not only in the ceiling-heavy evaluation regime. Recomputing over all four (Figure 4, Table 1), the organizing variable is position, not domain. The three interior-heavy datasets, including a second institutional corpus from a different country and instrument, are defined for every one of their cells; only our primary corpus, which piles against the ceiling, leaves about 42% of cells undefined, and its interior slice reproduces the interior behavior of the others. Positional informativeness is a property of where the data falls, and the bandwidth that carries it is interpretive resolution rather than computational uncertainty, so the display must encode it rather than assume it. Design, therefore, does not begin with a chart type, but with these three constraints, and asks what visual form keeps all three legible at once.

Table 2: Each property of the residual (Section 3) sets a display requirement, which in turn fixes an encoding choice.

Residual property	Requirement	Encoding choice
Qualifier, not verdict (R1)	Read with location	Position by rating; median and distribution adjacent (1, 3)
Distinct only where scale permits (R2)	Show where it adds value	Locate each item in its feasible band (1)
Informativeness is positional (R3)	Encode interpretive resolution	Band width and opacity (2)

4 Design Choices

A calibrated number is necessary but not sufficient. Read in isolation, the residual cannot show how far it resolves at a given rating level, nor the response shape that produced it, so an analyst who trusts a single value is most exposed exactly where the score is weakest, near the ceiling. The three properties of Section 3 say why no single number will do: the residual must be read with the median that locates it (R1), only in the part of the scale where it adds signal (R2), and with its positional informativeness shown rather than assumed (R3). The contribution is therefore the coordinated view that holds all three together, not the statistic in isolation.

We design for the people who act on these summaries, department chairs, program directors, and institutional research staff, who read rating data as part of decision support but do not necessarily reason in distributions [4]; the display externalizes the feasible-space reasoning rather than assuming the reader will supply it.

The three requirements translate into three design choices (Table 2), realized together in the view of Figure 5 and built on the principle that a summary is shown with, not instead of, its evidence [13, 24].

Showing feasible agreement, not raw agreement. To answer R1 and R2, each item is shown as a position within the band of agreement the scale permitted at its rating level (Figure 2, left). This is possibility-framing, in the lineage of quantile dotplots and hypothetical outcome plots, which externalize the space of outcomes rather than compress it to an interval [9, 14]. The residual becomes the visible gap between observed and expected rather than an opaque number, and because it is read against the feasible space, it is comparable across rating levels. Rating level and residual occupy the two position channels, which perception reads most accurately [5]. Without this framing, a reader sees only raw agreement, which restates the rating level and cannot be compared across it.

Encoding the positional informativeness of the calibrated score. To answer R3, the display shows how much the score can resolve at each position. The feasible bandwidth, which collapses toward the ceiling, carries this informativeness, and it cannot be left implicit. In the portfolio the width is mapped to opacity, so near-degenerate ceiling items recede while discriminating interior items stand out; the detail draws the band explicitly. This is uncertainty about the calibrated score itself, the meta-level, which the literature warns is routinely and consequentially omitted [12, 17]. Without it a reader takes a ceiling score at face value even though it carries no information, the failure mode that afflicts most high-rated items.

Keeping the calibrated summary adjacent to its distribution. To answer R1, an item is read at once by rating level and residual, with the median as its location. And because shape escapes any single number, a linked dot plot is one step away, honoring the commitment that distributional displays remain accessible [24, 30]; coordinated views bind the three into one reading [23]. The reading sequence is explicit (Figure 5): locate an item, read its agreement against what the scale allowed, judge how much that reading can

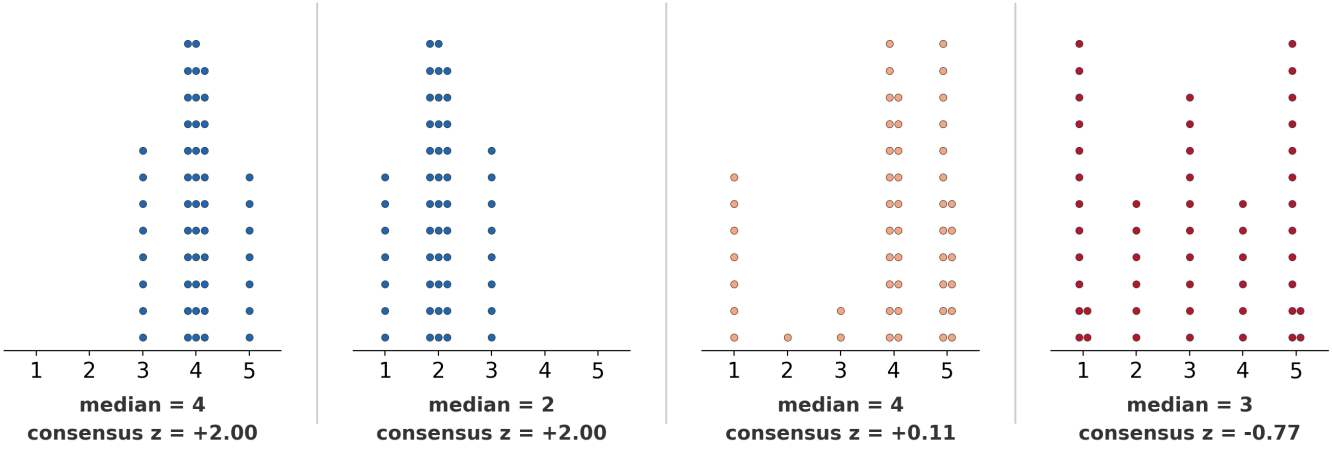


Figure 3: **The residual is a qualifier, read together with the median.** Four items, left to right: united and positive (median 4), united and negative (median 2), positive with a dissenting minority (median 4), and no consensus (median 3). The calibrated score alone is ambiguous: the same value is reassuring at a high median and alarming at a low one, and the first two share consensus $z = +2.0$ at opposite medians; the median keeps the overall lean honest under a dissenting minority. The median is the right partner because it resists a dissenting bloc and aligns with the median-anchored measure, whereas the mean and standard deviation move with both lean and dissent and cannot distinguish between these cases.

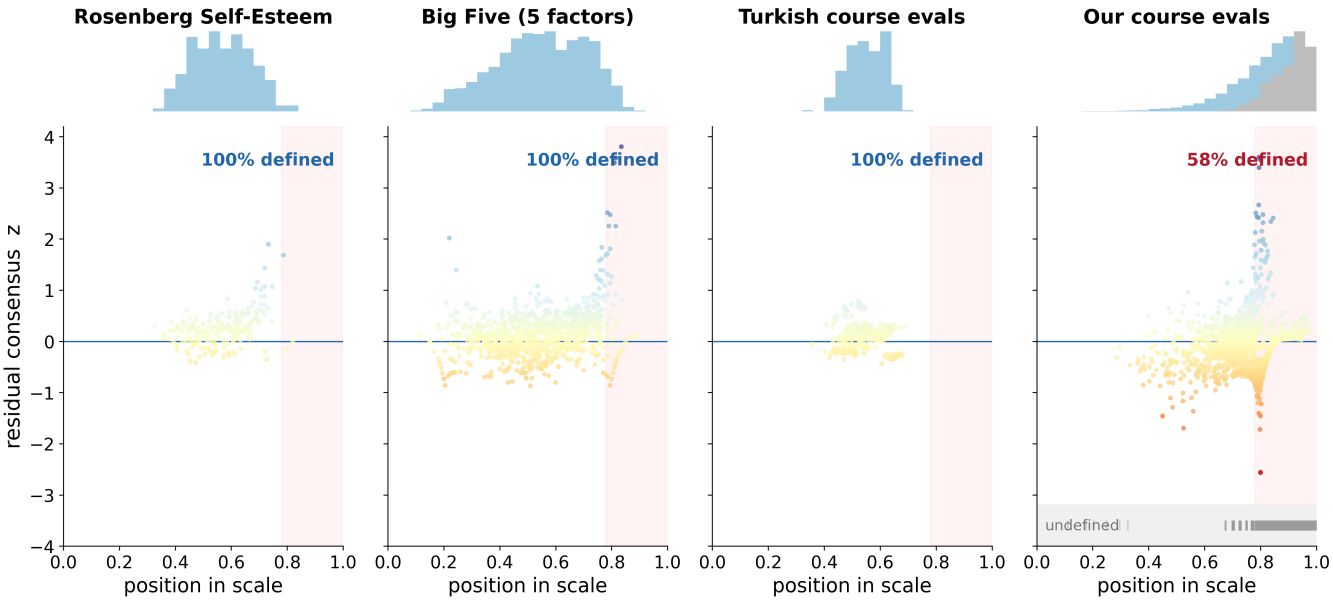


Figure 4: **Scale position, not domain, decides whether the residual is readable.** The two psychometric surveys are probe datasets, included to test the metric across the scale rather than as target domains; with a second institutional corpus, they are interior-heavy and defined everywhere, while only the primary corpus reaches the shaded ceiling zone, where the feasible space pinches shut and 42% of items are undefined. The collapse is positional, not domain-specific, and it is exactly what the portfolio down-weights through opacity and feasible bandwidth.

resolve, and open the distribution behind it. Without the adjacent distribution the calibrated scalar would hide the very structure it was meant to surface, since no single number separates all shapes at a given rating.

What the coordinated view adds. The view complements familiar displays rather than replacing them, and each does part of the job. A diverging stacked bar renders one item's full distribution and its balance around neutral, but it places the item against nothing, so it cannot say whether that agreement is surprising for the rating or how two items at different ratings compare [10, 28].

A dot plot shares that limitation; a box plot compresses to quartiles and discards the multi-modal shape, exactly the dissenting-subgroup and split structures that matter here. A table of means with error bars supports comparison across items but hides the distribution, and it inherits the ceiling effect that calibration removes [6]. The coordinated view is the combination the task demands: the portfolio positions every item on a calibrated axis so comparison is valid, opacity carries an informativeness signal a bar chart cannot express [7], and the linked distribution recovers the shape the scalar must drop. It is the pairing, not any one panel, that lets an analyst both make sense of a score and compare it across items; a table of means alone would report two items at the same rating as the same

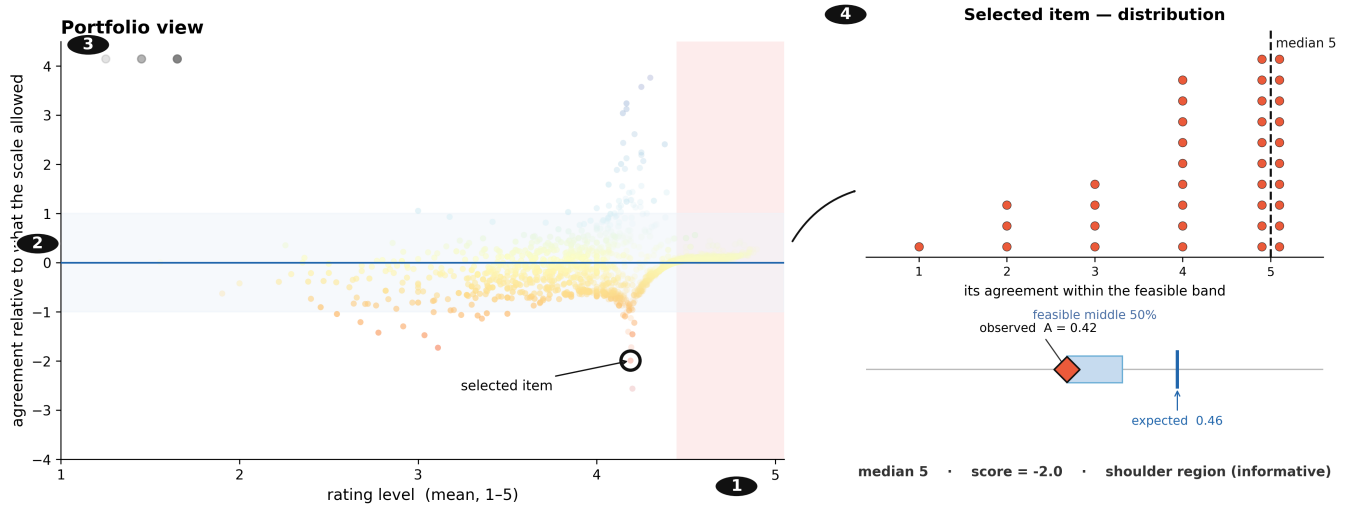


Figure 5: **One coordinated view makes agreement comparable and shows how informative each score is.** Left, a portfolio positioned by rating level and by agreement relative to what the scale allowed, with opacity encoding how much the score can resolve at that position and a shaded ceiling zone where the score degenerates. Right, a linked detail for a selected item: its distribution with the median marked, and its agreement within the feasible band, read as the gap from the expected value. The numbered badges trace a reading sequence: (1) locate an item by its rating level, (2) read its agreement against what the scale allowed there, (3) judge how much the score resolves from its opacity, and (4) open the distribution and median behind it. Mean and standard deviation alone cannot show that two items at the same rating differ in agreement, or that a score near the ceiling is less informative than one in the interior.

course.

The workflow the coordinated views support is traced in Figure 5. An administrator reviewing a department's evaluations first locates each course by its rating level and attends to the shoulder of the range, where the residual is defined and discriminating, letting the faint near-degenerate points against the ceiling fall away. She then reads each course's agreement against what the scale allowed there, and one course near a rating of 4.2 sits well below its band, a residual close to -1; its solid opacity, not the washed-out tone of a ceiling score, tells her the score is informative here. She opens its distribution and finds a strong majority at Excellent beside a small bloc at Poor, the dissenting subgroup of Figure 1. A table of means would have shown a 4.2 and moved on; the display shows a well-rated course with a subgroup whose experience diverged, worth following up. A neighbor at the same rating carries a large positive residual, genuine consensus with no dissent, and no action needed. Making sense of a single course and comparing two at the same rating are the two acts the display is built for, and both are steps this reviewer has just taken. Because the design is built on the residual's regional usefulness and positional informativeness, it also makes its own limits visible, which set the scope of the contribution.

5 Discussion

Summary statistics are representations. A summary is not a neutral readout of data; it is a representation, and a representation can introduce uncertainty of its own. Agreement on a bounded scale is one clear case: the map from responses to a score is many-to-one and its geometry changes across the scale, so the same number carries a different meaning depending on where it sits. Seeing summarization this way reframes a routine act, computing a statistic, as a step that can itself require uncertainty visualization, and therefore a target for visualization design rather than for statistics alone.

What residual consensus is and what it is not. A few clarifications sharpen what the calibration does. The reference against which agreement is measured, a uniform distribution over the response compositions that a given mean permits, is a deliberate and maximally noncommittal choice: it assumes nothing beyond the observed mean and count. It is not the only defensible reference, but the result depends on it only weakly. Under Dirichlet-multinomial alternatives, the population ordering stays stable, and only the extreme tail reshuffles, which is why we read the residual as a comparative ordering under a review budget rather than a fixed list of flagged items. The calibration is also not a re-description of variance. A median-anchored measure responds to the shape of a distribution, whereas variance responds to its spread, so on real data, the two correlate weakly and flag disjoint items; they behave as complementary instruments, each blind where the other sees (Section 3). Nor is the correction a form of normalization. Rescaling an agreement score to a common range would leave the difficulty untouched, because what changes across the scale is not the magnitude of the number but the set of agreements the scale can produce; comparability is restored by conditioning on that feasible set, not by any linear transform. What makes the calibrated position meaningful, in the end, is that the context it is read against is not chosen for convenience but is the mathematically constrained space of what the scale could have yielded.

Beyond agreement. The broader claim is that uncertainty can enter after data collection, at the point of summarization, and that such uncertainty is a visualization problem. Normalized scores, rates over small denominators, and composite indices are candidates for the same treatment: a summarizing transformation whose meaning varies across its domain, made legible by showing the feasible space behind the number. Residual consensus and its coordinated display are one concrete realization, and the agenda they point to is a visualization practice that communicates uncertainty introduced by transformation, not only uncertainty present in the data.

Scope and limitations. Several boundaries mark the current work, and each points toward a next step rather than qualifying the argument. The residual's distinctive signal is concentrated in the shoulder of the scale, since toward the interior it adds little to conventional dispersion and near the ceiling it flattens toward a constant; the display is built to show where it does and does not carry information, and to keep variance within reach for the firmly centered items, it reads better. Beyond the sensitivity we

already report, a fuller account of how the flagged set responds to the reference model and to the assumptions of exchangeable, evenly spaced responses would strengthen the method, and we leave that systematic study to future work. Our evidence rests on one institution's course evaluations alongside public psychometric data, which show that the metric's coverage generalizes across the scale; whether the same reading transfers to other domains and rating cultures is a question only further study can settle. Most of all, we have not yet put the display in front of the administrators it is meant to serve, and a controlled study with practitioners, testing both whether the encoding communicates as intended and how it shapes the decisions analysts reach, is the natural next step.

6 Conclusion and Future Work

On a bounded rating scale, agreement is not comparable across rating levels: the range a rating can occupy changes with the scale, so what a given score means is a matter of interpretation rather than estimation. Residual consensus answers it by reading agreement against the feasible space the scale allows, and the coordinated view makes that reading legible, paired with the median that locates it, marked with how far it can resolve, and kept beside the distribution that produced it. Together they let an analyst do what a mean and a spread cannot: separate genuine consensus from an artifact of the ceiling, and compare two items that a summary would call identical. The wider point outlasts the case. Uncertainty does not only inhabit the data; it is introduced when we summarize, and that summarization uncertainty is itself a target for visualization design.

We plan user studies as the immediate next step, in two parts. A controlled study with a larger sample would test the reasoning the display supports, whether readers judge agreement correctly across rating levels, discount scores near the ceiling appropriately, and compare two items at the same rating, and whether the coordinated view outperforms the calibrated number alone and a mean-and-distribution baseline. Beyond a single item, agreement can be calibrated at the level of a whole scale or factor, aggregating the feasible-space correction across related questions. The framing is not specific to course evaluations, and the same non-comparability afflicts other bounded scales, from product ratings to clinical and well-being instruments, which are natural sites to test whether the display generalizes to other domain contexts.

7 Acknowledgment

This work was supported in part by the National Science Foundation (NSF) grant 2312932.

References

[1] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973. 1

[2] J. Blair and M. G. Lacy. Statistics of ordinal variation. *Sociological Methods & Research*, 28(3):251–280, 2000. 2

[3] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. Overview and state-of-the-art of uncertainty visualization. In *Scientific visualization: Uncertainty, multifield, biomedical, and scalable visualization*, pp. 3–27. Springer, 2014. 2

[4] N. Boukhelifa, M.-E. Perrin, S. Huron, and J. Eagan. How data workers cope with uncertainty: A task characterisation study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 3645–3656, 2017. 2, 4

[5] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. 4

[6] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, 2014. 2, 5

[7] M. Correll, D. Moritz, and J. Heer. Value-suppressing uncertainty palettes. In *Proceedings of the CHI conference on Human Factors in Computing Systems*, pp. 1–11, 2018. 2, 5

[8] A. Dasgupta, M. Chen, and R. Kosara. Conceptualizing visual uncertainty in parallel coordinates. *Computer Graphics Forum*, 31(3pt2):1015–1024, 2012. 2

[9] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the CHI conference on human factors in computing systems*, pp. 1–12, 2018. 2, 4

[10] R. Heiberger and N. Robbins. Design of diverging stacked bar charts for likert scales and other applications. *Journal of Statistical Software*, 57:1–32, 2014. 2, 5

[11] J. M. Hofman, D. G. Goldstein, and J. Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020. 2

[12] J. Hullman. Why authors don't visualize uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):130–139, 2019. 2, 4

[13] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913, 2018. 2, 4

[14] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one*, 10(11):e0142444, 2015. 2, 4

[15] D. Kahneman and A. Tversky. Variants of uncertainty. *Cognition*, 11(2):143–157, 1982. 2

[16] A. Kale, M. Kay, and J. Hullman. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019. 2

[17] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):272–282, 2020. 2, 4

[18] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):892–902, 2018. 2

[19] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 5092–5103, 2016. 2

[20] R. K. Leik. A measure of ordinal consensus. *Pacific Sociological Review*, 9(2):85–90, 1966. 2, 3

[21] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005. 2

[22] J. Matejka and G. Fitzmaurice. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1290–1294, 2017. 1

[23] T. Munzner et al. *Visualization analysis & design*, vol. 1. CRC Press, 2015. 4

[24] L. Padilla, M. Kay, and J. Hullman. Uncertainty visualization. *Computational Statistics in Data Science. Wiley*, 2022. 2, 4

[25] L. M. Padilla, M. Powell, M. Kay, and J. Hullman. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, 11:579267, 2021. 2

[26] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997. 2

[27] R. H. Riffenburgh and P. A. Johnstone. Measuring agreement about ranked decision choices for a single subject. *The International Journal of Biostatistics*, 5(1), 2009. 2, 3

[28] N. B. Robbins, R. M. Heiberger, et al. Plotting likert and other rating scales. In *Proceedings of the Joint Statistical Meeting*, vol. 1. American Statistical Association, 2011. 2, 5

[29] M. Skeels, B. Lee, G. Smith, and G. Robertson. Revealing uncertainty for information visualization. pp. 376–379, 2008. 2

[30] L. South, D. Saffo, O. Vitek, C. Dunne, and M. A. Borkin. Effective use of likert scales in visualization evaluations: A systematic review. In *Computer Graphics Forum*, vol. 41, pp. 43–55. Wiley Online Library, 2022. 2, 4

[31] W. J. Tastle and M. J. Wierman. Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, 45(3):531–545, 2007. 2, 3

[32] C. Van der Eijk. Measuring agreement in ordered rating scales. *Quality and Quantity*, 35(3):325–341, 2001. 2, 3