Refinement: Measuring Informativeness of Ratings in the Absence of a Gold Standard

(Measuring Rating Refinement)

Sheridan Grant[*†], Marina Meilă[*], Elena Erosheva[*‡§], Carole Lee[¶]

March 23, 2023

This is the final version of the manuscript published by the British Journal of Mathematical and Statistical Psychology (Grant et al., 2022). It is no longer embargoed as it was published over one year ago.

The data that support the findings of this study are openly available in Figshare at https://doi.org/10.6084/m9.figshare.12728087. This work was partly supported by NSF grant #1759825 awarded to EE (PI) and CL (co-PI).

ORCIDs:

- Grant: 0000-0001-8952-2910
- Meilă: 0000-0002-3989-8853
- Erosheva: 0000-0003-2162-0017
- Lee: 0000-0002-5323-9205

[*]Department of Statistics, University of Washington, Seattle.
[†]Corresponding author, email slgstats@uw.edu
[‡]School of Social Work, University of Washington, Seattle.
[§]Center for Statistics and the Social Sciences, University of Washington, Seattle.
[¶]Department of Philosophy, University of Washington, Seattle.

CRediT Roles:

- Grant: Conceptualization, Methodology, Software, Validation, Formal Analysis, Writing (Original Draft), Writing (Review & Editing), Visualization
- Meilă: Conceptualization, Methodology, Validation, Formal Analysis, Writing (Original Draft), Writing (Review & Editing), Supervision, Project Administration, Funding Acquisition
- Erosheva: Conceptualization, Validation, Writing (Review & Editing), Project Administration, Funding Acquisition
- Lee: Conceptualization, Validation, Writing (Review & Editing), Project Administration, Funding Acquisition

Abstract

We propose a new metric for evaluating the informativeness of a set of ratings from a single rater on a given scale. Such evaluations are of interest when raters rate numerous comparable items on the same scale, as occurs in hiring, college admissions, and peer review. Our exposition takes the context of peer review, which involves uni- and multi-variate cardinal ratings. We draw on this context to motivate an information-theoretic measure of the *refinement* of a set of ratings—Entropic Refinement—as well as two secondary measures. A mathematical analysis of the three measures reveals that only the first, which captures the information content of the ratings, possesses properties appropriate to a refinement metric. Finally, we analyze refinement in real-world grant-review data, finding evidence that overall merit scores are more refined than criterion scores.

Keywords: Ratings, Entropy, Peer Review, Decision-Making

Acknowledgements

We thank the American Institute of Biological Sciences for the data used in this article, and particularly Stephen Gallo, AIBS Chief Scientist, for his responses to our queries regarding the AIBS review process and his feedback on an earlier draft of our paper.

1 Introduction

In this paper, we are interested in how experts communicate complex judgments via numerical ratings. Institutions often make decisions based on such ratings when the objectives of the decision process cannot be—or are too complex to be—mathematically formalized, when only human experts have the requisite decision-making knowledge, or when we want human value judgments to be expressed.[1] Humans typically make collective decisions via a *formal system* of rankings, ratings, or comparisons.[2]

Although researchers proposed statistical prediction as a replacement for clinical assessment decades ago (Meehl, 1954; Morera and Dawes, 2006), the more recent development of black-box machine learning algorithms has dramatically accelerated the switch from human to machine decision systems (Shortliffe and Sepúlveda, 2018; Athey, 2018). Because machine decisions can be formalized mathematically, they are analytically tractable. Specifically, the objective of the decision process can often be framed as an op-

[1]For example, Drury and Sinclair (1983) found that humans outperformed a machine in an industrial inspection task even though the machine was excellent at finding faults, because the machine was worse at determining the severity of the faults. The decision objective—fault severity—was difficult to formalize and the technology limited enough that human judgment was subtler and superior. Algorithms that aid judges in felony sentencing assess the risk of a defendant reoffending as well or better than humans can, yet judges routinely give younger defendants shorter sentences than recommended by algorithms "in line with a long-standing practice of treating youth as a mitigator in sentencing, due to lower perceived culpability" (Stevenson and Doleac, 2019).

[2]NIH grant proposal reviewers provide integer-scale ratings that are, after some discussion and possible revision, averaged (Staff, 2012). Maine began voting by ranked choice in 2018 (Staff, 2019a). Traditional first-past-the-post voting simply aggregates comparisons (Curtice, 2009).

timization problem in which the machine attempts to minimize predictive risk, a measure of how far from the true or optimal outcome a machine prediction/decision is. In contrast, humans often make decisions in contexts without a well-defined true outcome, which we will refer to going forward as a *gold standard*.

Current popular methods for analyzing human decision-making in the absence of a gold standard make comparisons to some other point of reference. For example, inter-rater reliability evaluates the extent to which one rater's ratings are replicated by a different rater. In this paper, we introduce the concept of *refinement*, an information-theoretic measure of the informativeness of a set of ratings from a single rater that makes no comparisons to a gold standard or other point of reference. Our exposition takes the context of peer review, in which human decision making is critical due to the lack of a gold standard by which to judge the predictive or external validity of peer review scoring practices (Bailar and Patterson, 1985; Feurer et al., 1994; Jayasinghe et al., 2001, 2003; Lauer and Nakamura, 2015; Lee and Moher, 2017; van Rooyen et al., 1999). Note that in peer review, prior research has shown that reliability may be a poor proxy for the normative credibility of review scores and review content (Bornmann et al., 2010; Lee et al., 2013; Hargens and Herting, 1990).

Different formal systems of human decision-making can lead to substantially different decisions (Langfeldt, 2001). The outcome of the popular vs. electoral college votes in the 2000 and 2016 U.S. presidential elections is a

prominent example. Comparisons between formal systems have typically focused on differences between rating scales. Schwarz et al. (1991) studied how changing global scale parameters without changing the internal structure of a scale affects raters' usage of the scale. In 1988, the National Institutes of Health (NIH) tested a move from a 1-to-5 decimal scale to a 1-to-5 scale with multiples of 0.5 (Green et al., 1989); more recently, NIH tested whether adding multiples of 0.5 to a 1–9 integer scoring system changed the distribution of average scores derived later in the process (Staff, 2019c). Neither study directly measured the utility of the decisions produced.

Attempts to circumvent the problem via a proxy gold standard have not found strong signals. In grant proposal peer review, this paper's motivating application and a textbook example of a cardinal rating system, Li and Agha (2015) found statistically significant gains in bibliometrics/productivity accruing from better NIH grant proposal scores. But Fang et al. (2016) and Lauer et al. (2015) find that on the whole these gains are practically modest, or even negligible. Note that the use of bibliometrics as a proxy for quality of scientific research is debated (Higginson and Munafò, 2016; Wang et al., 2017; Smaldino and McElreath, 2016; Lindner et al., 2018; Lindner and Nakamura, 2015).

Before introducing refinement in the peer review context, we briefly review rating systems, which are often used in contexts like peer review. In a rating system, raters score each item on a scale—possibly multiple scales, each representing a different aspect of the item. A *cardinal* scale's levels

have intrinsic numerical meaning via ratios or differencing (such as 0-100 essay grades), whereas Likert-type scales do not. Shah et al. (2014) found that pairwise comparisons are faster and, when aggregated, yield a more accurate ranking of the items than ratings. However, we restrict our attention here to cardinal scales, which yield fine-grained detail about the rated items in addition to an overall ranking. For example, in grant proposal peer review, ratings allow us to determine which applications meet a standard of quality rather than simply identifying the best ones. They may also facilitate providing applicants feedback that is more informative than simply their rank in a pool of anonymized applications. It is our goal to quantify the information produced by these complex systems.

1.1 Refinement

Refinement describes how finely a rater distinguishes between items of similar quality: do they give them all the same round score, or do they use small scale denominations to differentiate them? To what extent do the ratings imply an unambiguous ordering of the items? Refinement thus characterizes a set of scores from a single rater over multiple items, in contrast to reliability, which is a characteristic of ratings from multiple reviewers. As such, the rater is presumed to be interested in making distinctions among the items, meaning that the items must be comparable, such as grant proposals falling under the same round of review.

Refinement meets the immediate, practical need for a measure of the

utility of a set of ratings in the absence of a gold standard. At NIH, "there have been concerns that [the current 1-9 integer scale], which is functionally cut in half for the 50% of applications that are considered competitive, is not sufficient to express a study section's judgment of relative merit" (Nakamura, 2019). The Staff (2019c) study directly addressed this concern, but those analyses used aggregate ratings from multiple reviewers and did not consider individual reviewers' use of the scale. This study followed NIH's 2009 switch from a richer 1.0-to-5.0 single-decimal scale to the current 1-9 integer scale, a change motivated by the "compress[ed] score range" observed under the 1.0-to-5.0 scale which "effectively reduc[ed] the usefulness of scores for NIH funding decisions," as well as the difficulty of "[making] 41 reliable discriminations of application merit" (Staff, 2019b).

The Staff (2019c) study noted that "score compression and ties indicate that the review panel did not distinguish among the applications for impact and the lack of clear distinction among applications makes funding decisions more difficult, particularly when several applications receive identical scores and/or percentile ranks within the same study section." Thus, ambiguity of the ranking induced by the scores was a primary concern. The metrics used to measure this "score compression" were the frequency of ties at scores that were multiples of 10, and the percentiles of various common scores in the funding cutoff range—neither of which directly measure the extent to which the scores imply an unambiguous ranking of the applications or the quantity of information conveyed by the ratings. Refinement enables us to directly

assess the usefulness of a scale via the informativeness of ratings made on it.

We will adopt the language of the American Institute of Biological Sciences (AIBS) grant proposal peer review system, in which reviewers review applications/proposals—in general, *items*—providing scores/ratings on a set of *criteria* as well as an overall *merit* score. More precisely, we adopt the scenario in which a reviewer has several proposals to rate on a given scale. We shall measure the refinement of a set of ratings in a way that is sensitive to the fact that some reviewers can be assigned sets of proposals more similar in quality than others. Our measure will also account for the natural tendency of raters to prefer round ratings, which is explored at greater length in the next section.

Measuring refinement, or the degree of ranking disambiguity on a scale, will depend on the scale's allowing sufficiently fine comparisons between proposals close in value. In this paper, we design a refinement measure for decimal-based scales that admit multiples of 0.1 as ratings. The exposition employs the AIBS scale, which runs from 1.0 to 5.0, 1.0 being best, and admits a single decimal (Gallo et al., 2016). We denote the set of allowable scores $\mathbb{S}$, so that in this case $\mathbb{S} = \{1.0, 1.1, \ldots, 5.0\}$. Note that the refinement measures may depend on the scale used, and in this paper they will be tailored to the AIBS rating scale. We stress that refinement applies to rating systems generally, not just peer review at AIBS.

The next section lays out the refinement framework and our proposed primary measure of refinement—Entropic Refinement—with two additional

metrics briefly discussed for contrast. Section 3 compares mathematical and statistical properties of the metrics. Section 4 analyzes the refinement of the scores in a data set comprised of reviews of AIBS grant applications. The final section explains how refinement fits into the study of peer review, ratings, and decision-making more generally, and suggests directions for future work.

2 Measuring Refinement

In this section, we focus on a set of n univariate review scores from a single reviewer, denoted $[Y_1, \ldots, Y_n] = \mathbf{Y} \in \mathbb{S}^n$ (bold uppercase denotes random vectors, while standard uppercase denotes random variables). The n scores need not be unique. Hence, technically, $\mathbf{Y}$ is a *multiset*: a collection in which the elements need not be unique and order does not matter. For simplicity, however, we will continue to call it a set, with the understanding that the multiplicities are to be considered. The n scores correspond to n reviewed proposals, which are assumed to be in competition with one another—e.g., from the same round of review, in which the reviewer knows that their reviews will inform funding decisions.

When $\mathbf{Y}$ contains multiple scores with the same value, then the ordering of the respective items is not fully determined. A set of ratings $\mathbf{Y}$ is more refined when $\mathbf{Y}$ conveys more information about the relative ranking of the scored items. More specifically, the set of scores will distinguish finely between applications of similar quality, meaning there will be relatively fewer ties and

the ratings will imply a near-total ordering.

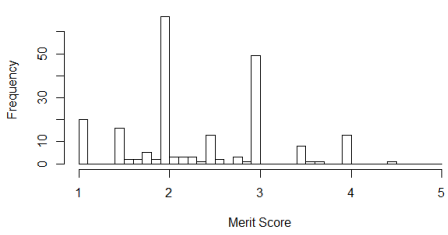
As we discuss in the next section, raters tend to provide round scores, inflating the likelihood of ties and decreasing refinement as we shall measure it. On the whole, finding ways to increase rating refinement should be useful to grant funding agencies and other institutions that use ratings to make decisions. However, we emphasize that refinement measures information, not the quality or accuracy of reviews. It may be that a peer reviewer, after careful consideration of a set of proposals, scores many of them equally. We therefore do not advocate blindly maximizing refinement; rather, refinement complements other techniques in the toolbox of ratings analysis.

2.1 Score Rounding

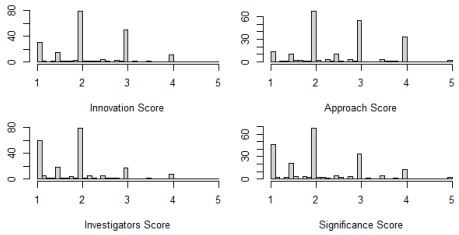
A rounding tendency has been shown to occur in multiple arenas, such as pricing (Lynn et al., 2013), price estimation (Simonsohn, 2013), age reporting (Grada, 2006), height reporting (Bopp and Faeh, 2008), cigarette-smoking reporting (Klesges et al., 1995), and length-ratio estimation (Plug, 1977). Relatedly, "heaping" describes how survey responses are often reported with an error that rounds the response to an integer number of units, e.g. "years married" or "income in thousands of dollars" (Bar and Lillard, 2012). Response set biases (Cunningham et al., 1977), such as extreme response bias (Erosheva et al., 2007), may also explain striking patterns in scoring such as the tendency to provide integer scores. For clarity, and since we do not wish to imply that providing a round score must involve an error, we use the term

rounding to refer to the tendency to provide scores that are multiples of 1 or 0.5 going forward.

The aggregated AIBS review data also provide strong evidence of rounding: AIBS reviewers use integer scores much more often than scores that are multiples of 0.5, which in turn are more frequent than other scores. This pattern is evident for merit scores and especially true for criterion scores; see Figure 1 (Section 4.3 also lends support to this claim). There is thus ample evidence that AIBS reviewers gravitate towards rounder scores. Yet providing a less ambiguous comparison of the proposals requires resistance to this pull. Our refinement metric shall measure the extent to which reviewers do so.



(a) Merit score frequencies.



(b) Criterion score frequencies.

Figure 1: Histograms of scores from the AIBS data, all 216 reviews from all 26 reviewers. Rounder score levels (multiples of 0.5, particularly multiples of 1) are clearly preferred.

2.2 Entropic Refinement

We now introduce an entropy-based refinement metric that explicitly measures the extent to which reviewers resist the rounding tendency, via the de-

crease in entropy induced by rounding the scores. Entropy is an information-theoretic measure of how unpredictable the data generated from a probability distribution is. For a probability mass function p on k elements defined by $p_1, \ldots, p_k$, the Shannon entropy of p is

$$H(p) = -\sum_{i=1}^{k} p_k \log p_k$$

(Cover and Thomas, 2012).[3] In our application, p is the empirical distribution on $\mathbb{S}$ induced by $\mathbf{Y}$, and we will use the shorthand $H(\mathbf{Y})$ instead for clarity. Let $R_t(y)$, $t \in \{0.5, 1\}$ be a rounding function that rounds y to the nearest multiple of 0.5 or 1, and extend it to vectors and matrices in the natural way. Define $\mathbb{S}_t \equiv R_t(\mathbb{S})$ the set of possible scores after rounding to level t; here, for example, $\mathbb{S}_1 = \{1, 2, 3, 4, 5\}$.

Entropic Refinement is our proposed refinement metric and is defined as the decrease in entropy induced by rounding:

$$r_E(\mathbf{Y}; t) \equiv H(\mathbf{Y}) - H(R_t(\mathbf{Y})). \tag{1}$$

Moving forward, we will assume integer rounding ($t = 1$) unless otherwise specified and will drop t from the notation unless it is needed for clarity.

Note that, because rounding is a form of quantization and quantization can only reduce entropy (Cover and Thomas, 2012), Entropic Refinement is

[3] We take the log base e; this is merely an arbitrary choice of scale, and base e makes certain mathematical manipulations of H simpler.

non-negative. Entropic Refinement will tend to be higher when scores are not disproportionately round, as is observed in Figure 1, but rather spread evenly around round score levels (see Section 3.1).

Entropic Refinement aligns with the following behavioral reviewing example. Suppose that reviewers first choose a plausible score from a coarser subset of the available scores, such as the integers or multiples of 0.5. After this initial scoring, there are likely to be numerous ties between proposals' scores, which could prompt reviewers to then adjust the scores by small amounts based on further evaluation or by comparison to previously-rated proposals. The greater the extent of these adjustments, the more information the scores provide, and the larger Entropic Refinement grows. In the context of this example, Entropic Refinement (1) can be interpreted as the increase in entropy induced by adjusting the initially rounded scores.

We formally analyze the properties of Entropic Refinement in Section 3. First, however, we briefly discuss two alternate approaches to measuring refinement which, while intuitive and appealing at first glance, will be shown to be inadequate. They are useful as bases of comparison when considering the properties of Entropic Refinement.

2.3 Alternative Refinement Metrics

The following alternate metrics, Fractional and Tiebreak Refinement, were each constructed to target a specific aspect of refinement: the tendency to avoid rounding and the tendency to break ties. We do not advocate for these

measures because, as we illustrate in what follows, each ignores an important facet of refinement.

2.3.1 Fractional Refinement

Given a decimal-valued scale like that used by AIBS, we may assert that utilizing decimal values beyond just multiples of 1.0—or even 0.5—conveys more refinement.

Taking Y_i to be the ith score in $\mathbf{Y}$, let $n_{0.5}$ be the number of scores that are multiples of 0.5 but not 1 and $n_{0.1}$ be the number of scores that are not a multiple of 0.5 (or 1). That is, $n_{0.5} \equiv \sum_{i=1}^{n} \mathbf{1}[Y_i \pmod 1 \neq 0]\mathbf{1}[Y_i \pmod{0.5} = 0]$ and $n_{0.1} \equiv \sum_{i=1}^{n} \mathbf{1}[Y_i \pmod{0.5} \neq 0]$. Let $w \in (0, 1)$ be a weight parameter. Then we define the *Fractional Refinement* of $\mathbf{Y}$ to be

$$r_F(\mathbf{Y}; w) \equiv \frac{1}{n}(n_{0.1} + w n_{0.5}). \tag{2}$$

Thus, Fractional Refinement is a linear combination of the frequencies of the different types of scores, where rounder ratings receive less weight (integer scores receive zero weight). Fractional Refinement is a straightforward way of determining whether reviewers are utilizing all the types of levels the scale provides.

However, Fractional Refinement does not directly measure informativeness. For example, a set of identical scores $\mathbf{Y} = [3.2, \ldots, 3.2]$ has maximal Fractional Refinement, but does not help us distinguish between the propos-

als at all.

In contrast, our next secondary refinement metric—Tiebreak Refinement—directly measures the extent to which applications are ranked unambiguously.

2.3.2 Tiebreak Refinement

A refined set of ratings conveys small differences between applications' perceived quality and will thus contain relatively more small differences between scores than ties.[4] We think of these small differences as potential evidence that reviewers recognize when applications are of similar quality but then break rating ties in order to indicate the applications' relative ranking. This motivates the Tiebreak Refinement metric.

Let $Y_{(i)}$ be the ith order statistic of $\mathbf{Y}$, with ties broken arbitrarily and $\mathbf{D}(\mathbf{Y}) \equiv \left\{Y_{(i+1)} - Y_{(i)} : i \in [n-1]\right\}$ be the multiset of distances between consecutive scores. Then

$$z(\mathbf{Y}) = |\{x \in \mathbf{D}(\mathbf{Y}) : x = 0\}|$$
$$l(\mathbf{Y}, c) = |\{x \in \mathbf{D}(\mathbf{Y}) : 0 < x \leq c\}|$$

define the "zero" sorted distances (ties) and the "little" sorted distances for some $c < 1$ (just z and l when context is clear). We then define *Tiebreak*

[4] Ties may not reflect true evaluative equality when n is not sufficiently smaller than the number of levels on the rating scale, an issue that arises for a small subset of the reviewers in our AIBS application and that we also address in Section 3.2.

Refinement as the fraction of sorted distances less than c that are nonzero:

$$r_T(\mathbf{Y}; c) \equiv \frac{l}{z + l}. \tag{3}$$

If $z = 0$ and $l \neq 0$, then $r_T = 1$; if $l = 0$ and $z \neq 0$, then $r_T = 0$. If both l and z are zero, then we set $r_T = 1$ because all sorted distances are large and there are no ties. For $n = 1$, because there are no sorted distances, r_T is undefined.

The choice of c is application-dependent; for the AIBS scale, we recommend $c < 0.5$, and in our Section 4 application we choose $c = 0.2$ so that every score is at most a "little" distance from exactly one multiple of 0.5.

Tiebreak Refinement is the foil of Fractional Refinement: it directly measures ties, but in no way accounts for rounding and the structure of the scale. Consider two sets of scores $\mathbf{Y} = [1, 1.5, 2, 2, 2.5]$ and $\mathbf{Y}' \equiv \mathbf{Y} + 0.1 = [1.1, 1.6, 2.1, 2.1, 2.6]$. Then $r_T(\mathbf{Y}; 0.5) = r_T(\mathbf{Y}'; 0.5) = 3/4$, but rounding may have taken place for the $\mathbf{Y}$ scores, while it certainly has not for $\mathbf{Y}'$. What *is* clear for both sets of scores is the rank order of the proposals.

3 Properties of Refinement Metrics

This section derives mathematical properties of Entropic Refinement and compares them to those of Fractional and Tiebreak Refinement. We also highlight how these properties should inform applications and interpretations

of refinement. Properties are summarized in Table 1.

Property (Section)	Entropic Refinement	Fractional Refinement	Tiebreak Refinement
Decomposition (3.1)	Basins of Attraction	Trivial	None
Range (3.2)	$[0, \log \max_{s \in \mathbb{S}}\{\lvert B(s)\rvert\}]$	$[0, 1]$	$\left[0, \frac{\lvert\mathbb{S}\rvert - 1}{n-1}\right]$
Large-n (3.3)	Normalized	Normalized	$\lim_{n\to\infty} r_T(\mathbf{Y}; c) = 0$

Table 1: Summary of properties of the three refinement metrics. Associated sections provide full explanations.

3.1 Decomposition

Rounding is a common operation on scores that also naturally induces a partitioning of the scale $\mathbb{S}$. We now demonstrate how Entropic Refinement can be decomposed in terms of this partitioning.

For every $s \in \mathbb{S}_t$, define its *basin of attraction* $B_t(s) \equiv \left\{R_t^{-1}(s)\right\}$ to be the set of scores that yield s when rounded to level t. Clearly, the sets $B_t(s)$ for $s \in \mathbb{S}_t$ partition $\mathbb{S}$. These basins need not all be the same size: for AIBS, $|B(1)| = |\{1.0, 1.1, 1.2, 1.3, 1.4\}| = 5$, $|B(5)| = |\{4.5, 4.6, 4.7, 4.8, 4.9, 5.0\}| = 6$, and $|B(s)| = 10$ for $s \in \{2, 3, 4\}$.

Again for every $s \in \mathbb{S}_t$, let $\mathbf{Y}_{s,t} \equiv \{y \in \mathbf{Y} \cap B_t(s)\}$. Also recall that p is the empirical probability mass function on $\mathbf{Y}$, so that $p(B(s))$ is the fraction of the observed scores that lie in $B(s)$. We then have the following decomposition result, whose proof can be found in Appendix A.1:

Proposition 1. *For any score vector* $\mathbf{Y}$ *and rounding level* t*, Entropic Re-*

finement is a weighted average of entropies over rounding basins:

$$r_E(\mathbf{Y}; t) = \sum_{s \in \mathbb{S}_t} p(B_t(s)) H(\mathbf{Y}_{s,t}). \tag{4}$$

The weights $p(B_t(s))$ are the fractions of the observed scores in each basin, and the entropies $H(\mathbf{Y}_{s,t})$ represent the quantity of information conveyed by the scores within each basin. No disambiguation between scores that are in the same basin leads to zero refinement, whereas breaking ties within a basin (disambiguation) leads to increased within-basin entropy and increased Entropic Refinement. Scores that are close in that they are in the same basin interact with one another in determining Entropic Refinement, but not with scores lying in other basins.

We now briefly analyze decomposition properties for Fractional and Tiebreak Refinement. For Fractional Refinement

$$r_F(\mathbf{Y}; w) = \frac{1}{n} \sum_{i=1}^{n} r_F(Y_i; w)$$

by linearity, since r_F is simply a weighted average. Such trivial decomposability is an undesirable property, because it means that r_F fails to take into account relationships between the scores.

Tiebreak Refinement cannot be decomposed at all: hiding the value of a single Y_i makes $Y_{(j)}$ indeterminate for all j, so Tiebreak Refinement of proper subsets of the observed scores $\mathbf{Y}$ cannot fully inform us of the Tiebreak

Refinement of the full set. While Tiebreak Refinement captures dependence among the scores, it ignores rounding and does not respect the local structure of the scale, as Entropic Refinement does via basins of attraction.

3.2 Extrema and Range

We now turn to upper and lower bounds for each refinement metric. These results show how different scales may not be easily comparable in terms of refinement. We recommend only comparing refinement metrics derived from the same scale.

Proposition 2. *The Entropic Refinement r_E takes values between* 0 *and* $\log \max_{s \in \mathbb{S}_t} |B(s)|$*, attaining its minimum when there is only one unique observed score value in any basin of attraction and its maximum when scores are only located in the maximally sized basins of attraction, and are uniformly distributed within each such basin.*

See Appendix A.2 for proof. To see how this maximum is attained, and that it depends on the rounding level t, consider the following 3 sets of scores: $\mathbf{Y}_A = \{1.0, 1.1, \ldots, 5.0\}$; $\mathbf{Y}_B = \{1.5, 1.6, \ldots, 4.4\}$; and $\mathbf{Y}_C = \{1.3, 1.4, \ldots, 4.7\}$. $\mathbf{Y}_A$ is uniform over the entire scale, $\mathbf{Y}_B$ over the maximum-size basins for $t = 1$, and $\mathbf{Y}_C$ over the maximum-size basins for

$t = 0.5$. Hence,

$$\begin{aligned} r_E(\mathbf{Y}_A; t = 1) &= \frac{5}{41}\log(5) + 3 \times \frac{10}{41}\log(10) + \frac{6}{41}\log(6) \\ &\approx 2.14 \\ &< \log(10) = r_E(\mathbf{Y}_B; t = 1). \end{aligned}$$

While $\mathbf{Y}_B$ maximizes $r_E(\mathbf{Y})$ for $t = 1$, $\mathbf{Y}_C$ does for $t = 0.5$:

$$\begin{aligned} r_E(\mathbf{Y}_C; t = 0.5)) &= 7 \times \frac{1}{7}\log(5) \\ &\approx 1.61 \\ &> 1.50 = r_E(\mathbf{Y}_B; t = 0.5). \end{aligned}$$

This property holds not only for different levels of rounding but also for different scales. Consider $2\mathbb{S}$, a 2–10 scale that admits only multiples of 0.2. The only difference between this scale and the AIBS scale is a factor of 2, but refinement metrics calculated from each will be incomparable because, for example, $2\mathbb{S}$ contains nine integers instead of five.

Fractional Refinement achieves its minimum of zero when $\mathbf{Y}$ is integral and its maximum of 1 when $\mathbf{Y}$ does not contain multiples of 0.5.

Tiebreak Refinement achieves its minimum of zero if and only if $l = 0$ and $z > 0$, for example when only integer scores are present and there is at least one tie. It achieves its maximum of 1 whenever there are no ties, i.e., $z = 0$. However, when $n > |\mathbb{S}|$, $z > 0$ necessarily. In this case, the maximum

Tiebreak Refinement for a given $n > |\mathbb{S}|$ is $\frac{|\mathbb{S}|-1}{n-1}$, which occurs whenever every level of the scale is utilized, i.e. when $\mathbb{S} \subseteq \mathbf{Y}$.

3.3 Large-n Behavior

Here, we consider the dependence on n of the Entropic Refinement r_E. For $n = 1$ and any $\mathbf{Y}$, $r_E(\mathbf{Y}) = 0$. For $n = 2$, r_E is zero when the two scores are identical or in different basins of attractions, and $\log(2)$ if the two scores are different but in the same basin of attraction; but $\log(2)$ is still much smaller than the maximum r_E on the AIBS scale for arbitrary n, which is $\log(10)$. It is clear that, for small values of n, the dependence of r_E on n is strong. Therefore, when sample sizes are modest, comparisons of refinement statistics must be stratified by n.

However, the behavior of r_E as $n \to \infty$ tells another story. The following analysis draws a contrast between Entropic and Fractional Refinement on one side and Tiebreak Refinement on the other. For both Fractional and Entropic Refinement, given any set of scores $\mathbf{Y} \in \mathbb{S}^n$, we can construct an infinite sequence of sets of scores $\mathbf{Y} \in \mathbb{S}^n, [\mathbf{Y}, \mathbf{Y}] \in \mathbb{S}^{2n}, [\mathbf{Y}, \mathbf{Y}, \mathbf{Y}] \in \mathbb{S}^{3n}, \ldots$ (repeat each score in $\mathbf{Y}$ once, twice, etc.) such that refinement is constant within the sequence. We call a refinement metric for which such a sequence exists for any $\mathbf{Y} \in \mathbb{S}^n$ *sample size-normalized* for large n.

This is not the case, however, for Tiebreak Refinement. It follows directly from the fact that $r_T \leq \frac{|\mathbb{S}|-1}{n-1}$ (Section 3.2) that $\lim_{n\to\infty} r_T = 0$, and hence such an infinite sequence with constant refinement does not exist. The differ-

entiating factor is that Fractional and Entropic Refinement are functions only of the empirical distribution on $\mathbf{Y}$—given the empirical distribution, they are independent of the sample size n and the observed scores themselves—while Tiebreak Refinement depends on the scores themselves. We argue that the informativeness of a set of ratings should only depend on its distribution and not tend to zero as more ratings are given.

3.4 Multivariate Extensions

In some peer review systems, such as those at AIBS and NIH, reviewers provide C *criterion scores* $X^1, \ldots, X^C$ in addition to the merit score Y. The criterion scores are intended to be preliminary to the merit score, as well as provide more detailed feedback to applicants on various aspects of their application. They are rated on the same scale $\mathbb{S}$ as the merit score Y, which is used to determine proposal funding. Here we present methods for assessing the refinement of multivariate scores such as criterion scores.

Let C be the number of criteria in the reviewing system at hand and assume (as is the case for AIBS) that each criterion is measured on the same scale as the merit score Y. Let $X^1, \ldots, X^C$ denote the individual criterion scores with $\mathbf{X} \in \mathbb{S}^C$ the vector of criterion scores. The most immediate way of measuring multivariate refinement is to average over the C dimensions:

abusing notation slightly, we define

$$r^{avg}(\mathbf{X}) \equiv \frac{1}{C} \sum_{k \in [C]} r\left(\mathbf{X}^k\right)$$

for a given metric r. For Fractional and Tiebreak Refinement, this average is the only clear choice. However, for Entropic Refinement, we can also consider the entropy of the empirical joint p.m.f. of the criterion scores,

$$r_E^{joint}(\mathbf{X}) \equiv H(\mathbf{X}) - H(R(\mathbf{X})).$$

This extension is fundamentally different than the average over individual criteria. The next proposition clarifies the relationship between the two.

Proposition 3. *For any* $\mathbf{X} \in \mathbb{S}^{C \times n}$ *and a given* t*,*

$$\frac{1}{C} \sum_{k \in [C]} r\left(\mathbf{X}^k\right) = r_E^{avg} \leq r_E^{joint} \leq C r_E^{avg} = \sum_{k \in [C]} r\left(\mathbf{X}^k\right) \tag{5}$$

See Appendix A.3 for proof. The left-hand inequality becomes equality when, for example, all criterion scores are identical for each proposal. The right-hand inequality becomes equality when $X^1, \ldots, X^C$ are mutually independent with respect to their joint empirical distribution. In practice, the criterion scores are likely to be correlated to some extent, and both inequalities will be strict. Thus scaling r_E^{joint} so as to be comparable between vectors of scores with different values of C is impractical. For this reason, we use av-

Statistic	Value
Number of reviewers N	26
Number of applications	72
Total number of reviews $\sum_{i=1}^{26} n_i$	216 $(= 3 \times 72)$

Table 2: AIBS data set summary statistics.

erage multivariate refinement in Section 4's application to AIBS peer review data.

4 Refinement in AIBS Grant Proposal Peer Review Scores

Using the refinement metrics introduced above, we analyze the scoring behavior of reviewers at the biomedical science grant agency AIBS (American Institute of Biological Sciences). The University of Washington's Institutional Review Board confirmed that this study did not directly involve human subjects.

4.1 AIBS Review Data

The AIBS data set consists of review scores of 72 grant applications, all from the same round of review, reviewed by AIBS through an intramural collaborative biomedical research funding program for the biomedical sciences (Gallo, 2021). For each application, exactly three reviewers provide four criterion scores—Innovation, Approach, Investigator, and Significance—on

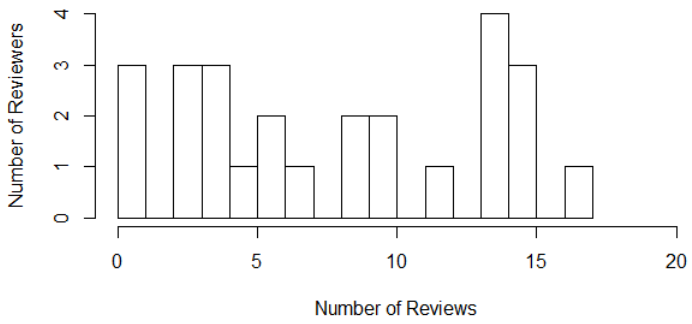


Figure 2: Histogram of the number of proposals reviewed by each reviewer.

a 1-to-5 scale in single-decimal (0.1) increments, where 1 is best and 5 is worst. AIBS reviewers also supply a merit score that attempts to capture the quality of the entire application, rather than just an aspect of it as the criterion scores do. Funding decisions are made largely on the basis of these merit scores.[5]

Let $N = 26$ refer to the number of reviewers and n_i be the number of reviews performed by the ith reviewer. Table 2 displays summary statistics for the data set, and Figure 2 displays a histogram of the number of scores given by each reviewer, i.e. a histogram of $\{n_i : i \in [N]\}$.

4.2 Testing Refinement Hypotheses

We exemplify the use of refinement for the study of reviewer behavior by testing a hypothesis regarding refinement that we believed *a priori* to hold for AIBS and more broadly across peer review systems.

[5]The criterion scores, in addition to other factors such as the topic of the proposed research, can also play roles in these decisions.

Hypothesis 1. ***Merit score refinement is higher than criterion score refinement:*** For AIBS as well as other funding agencies, the merit score (or equivalent) is the primary score used in funding decisions. Reviewers may therefore attempt to finely distinguish between similar-quality applications via merit scores, while allowing for more ties in criterion scores, which may instead be considered a mechanism for providing detailed feedback to applicants. If this were so, we would expect merit scores to display more refinement than criterion scores. The null hypothesis we test is that merit score refinement is less than or equal to criterion score refinement.

We utilize average criterion score refinement, r_E^{avg}, when computing Entropic Refinement as the criterion scores and merit score are of different dimensions (see Section 3.4). We use the paired t-test to test the null hypothesis that merit score refinement is no greater than the average criterion score refinement: $r(\mathbf{Y}) \leq r^{avg}(\mathbf{X})$. We use the Wilcoxon signed-rank test (Wilcoxon, 1946) to test the null hypothesis that $\mathrm{P}\left(r(\mathbf{Y}) > r^{avg}(\mathbf{X})\right) \leq 0.5$,[6] with the alternative hypothesis being that $\mathrm{P}\left(r(\mathbf{Y}) > r^{avg}(\mathbf{X})\right) > 0.5$. The sample size for both tests is the number of reviewers $N = 26$.

Both tests are paired, so that merit and criterion score refinement are always compared on an individual level. However, we do not stratify the tests by the number of reviews completed. While this does not harm the Type I error rates of the tests, it means that reviewers with larger n_i are

[6] If one randomly samples a reviewer and associated review scores from their respective hypothetical populations, then there is at most a 50% probability that the merit score refinement will be greater than the average criterion score refinement.

weighted more heavily in the tests. We believe this is appropriate, given that these reviewers completed more reviews, but we do not claim that our testing strategy is fully efficient—developing maximally efficient tests for refinement is a new problem entirely.

Both of these tests operate under the assumption of independent observations, i.e. that the refinement of the scores from one reviewer is independent of the refinement of the scores from a different reviewer.[7] We assume that two sets of scores are independent when the underlying proposals reviewed do not overlap, so that refinement of one reviewer's scores does not depend on other reviewers' scores except possibly when reviewers review the same proposal. Per AIBS, "online discussion [among reviewers] was limited and most scoring did not change [after online discussion]," so the only plausible reason for dependencies between refinement statistics is overlap in the underlying proposals reviewed. In our data, the rate of overlapping reviewer assignment is small (see Figure 3), so we believe the assumption of independent observations made by these tests is reasonable.

4.3 Results

We compute Entropic Refinement for the merit and each of the four criterion scores individually, for each of the 26 reviewers, with rounding either to the

[7] We make no assumptions regarding the dependencies among the scores from a single reviewer. Since the unit of observation for testing this hypothesis is a refinement statistic for *all* scores from a given reviewer, these dependencies are not material to this hypothesis test.

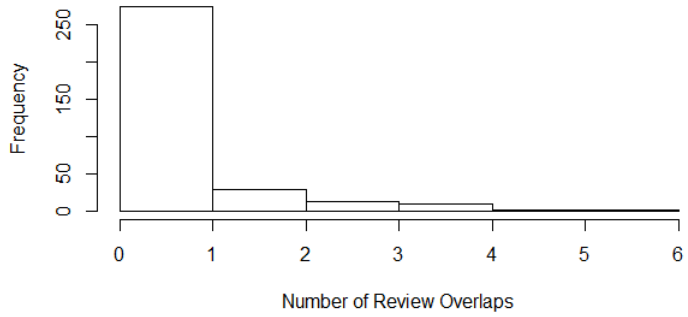


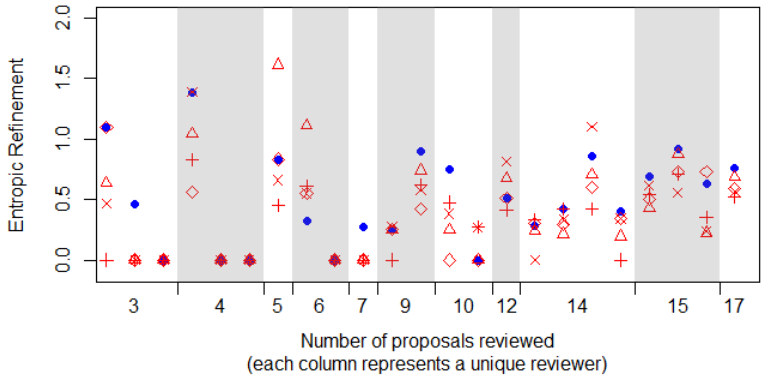
Figure 3: Histogram of the number of overlapping proposal assignments for each pair of reviewers. There were two pairs for which this quantity was 6, the maximum, both of which involved reviewers of 14 or 15 proposals.

nearest integer or the nearest multiple of 0.5. First, we plot r_E in Figure 4 for merit and the criteria.

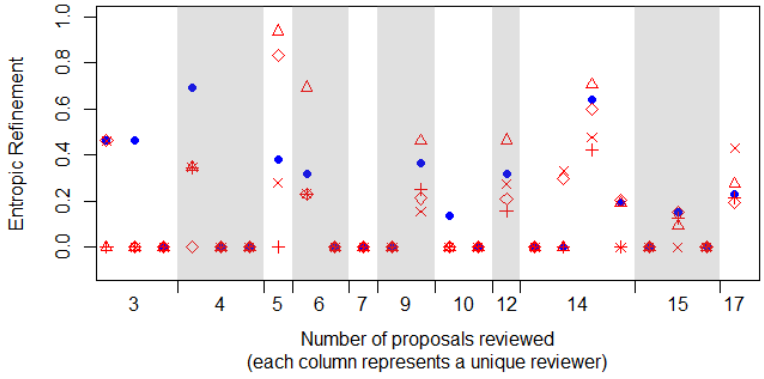
We also compare the entirety of the empirical distributions of merit and criteria Entropic Refinement for the $N = 26$ reviewers, which are displayed in Figure 5. The empirical distribution of merit refinement stochastically dominates that of criteria refinement in the $t = 1$ case, and nearly does so in the $t = 0.5$ case.

Table 3, below, illustrates the results of both hypothesis tests—paired t-test and Wilcoxon signed-rank—applied to the AIBS merit and criterion refinement. Each test was specified to be one-sided, with the alternative that merit refinement is greater than criterion refinement. For reference, tests using the Fractional and Tiebreak metrics were included as well.

For all tests, the evidence suggests that merit scores display greater refinement. p-values for Fractional Refinement are significant at the 0.005 level,

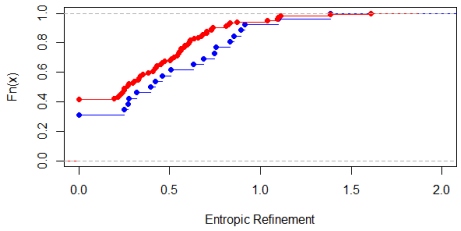


(a) Entropic Refinement with rounding to the nearest $t = 1$.

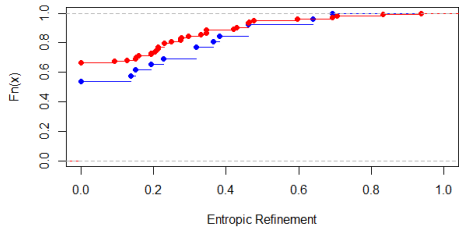


(b) Entropic Refinement with rounding to the nearest $t = 0.5$

Figure 4: Entropic Refinement for the AIBS reviewers, with merit score refinement in blue and criterion score refinement in red ($\triangle$ = Innovation, $+$ = Approach, $\times$ = Investigators, $\diamond$ = Significance). Individuals columns of points are unique to reviewers, and represent their ranking in terms of number of proposals reviewed in order from least to greatest. The number of proposals reviewed is indicated on the x axis and by the alternated shaded/unshaded columns. Reviewers with $n = 1$ proposal reviewed have $r_E = 0$ by default and are not displayed.



(a) Entropic Refinement with rounding to the nearest $t = 1$.



(b) Entropic Refinement with rounding to the nearest $t = 0.5$.

Figure 5: Entropic Refinement empirical CDFs for the AIBS reviewers, with merit score refinements in blue and criterion score refinements in red.

Refinement Metric	Parameter	t-test mean dif. (p-val)	Wilcoxon p-val
Entropic	$t = 1$	0.12 (0.004)	0.003
	$t = 0.5$	0.049 (0.04)	0.05
Fractional	$w = 1$	0.10 (0.001)	0.001
Tiebreak	$c = 0.2$	0.11 (0.01)	0.02

Table 3: Test statistics and p-values for the three types of refinement.

meaning that rounder scores are used significantly less often for the merit score than for the criterion scores on the whole. p-values for Tiebreak Refinement are only significant at the 0.05 level, suggesting that merit score ranking ambiguity is higher for the criterion scores than for the merit scores. Finally, for Entropic Refinement, p-values are significant at the 0.005 level when rounding to the nearest integer but barely significant at the 0.05 level for rounding to the nearest multiple of 0.5. There is thus strong evidence for merit scores being more informative within integer rounding basins, but weaker evidence for a difference in informativeness within half-integer rounding basins. All in all, the evidence is moderately strong that AIBS merit scores, which help determine proposal funding decisions, are more refined

than criterion scores, even in our fairly small sample of reviews.

5 Discussion

In this paper, we articulated the concept of refinement as a novel way of quantifying the informativeness of human ratings that is particularly useful in the absence of a gold standard. We introduced an information-theoretic metric for measuring it—Entropic Refinement—and examined refinement for a set of reviews of grant proposals submitted to AIBS.

Entropic Refinement captures disambiguation through the difference in entropy before and after rounding the scores. As Proposition 1 demonstrates, this is equivalent to measuring the entropy—information content—of the scores within basins of attraction, and taking a probability-weighted average. While Entropic Refinement is more complex than the two simpler metrics introduced in the paper, its decomposability property, sensible asymptotic behavior, and behavioral motivation make Entropic Refinement our recommended metric.

Refinement measures informativeness of reviews, remaining agnostic to differences in their underlying quality. One must still bear these differences in mind when interpreting refinement, however, as we generally expect refinement to be higher when the differences between the proposals reviewed are larger. There are various ways of controlling for differing proposals and/or reviewers. When all reviewers review all proposals, differences between re-

viewers' refinement cannot be attributed to differences between proposals. If reviewers have been randomized into groups whose peer review processes have been altered in various ways, then differences in refinement between groups cannot be attributed to systematic differences in proposal quality, even though the proposals reviewed in the groups may differ. Finally, in our study of AIBS data, we use a matched design to compare differences in refinement between types of scores: the proposals, reviewers, and assignment of proposals to reviewers are identical in the two groups being compared.

Because refinement measures the degree of disambiguation and informativeness of a reviewer's scores without comparison to some external baseline, such as a different reviewer's scores, it is distinct from inter-rater reliability metrics and particularly well suited to the no-gold-standard paradigm, as typically holds in peer review (Bailar and Patterson, 1985; Feurer et al., 1994; Jayasinghe et al., 2001, 2003; Lauer and Nakamura, 2015; Lee and Moher, 2017; van Rooyen et al., 1999).

Refinement may also provide important context in conjunction with inter-rater reliability. For example, when reliability is low, low refinement across reviewers means that there is neither consensus nor abundant information about the relative merits of the proposals. High refinement paired with low reliability, however, suggests that while they may not agree, reviewers are effectively disambiguating the proposals they rate—potential evidence of the use of a variety of evaluative perspectives. High reliability with low refinement implies the opposite, and may indicate that the scale at hand

is insufficiently fine-grained for reviewers to assert their unique perspectives (this may or may not be desirable, depending on the setting). Finally, high reliability and high refinement—likely a rare outcome[8]—would imply that consensus is not merely the product of a coarse scale or heavily rounded ratings.

In psychology, ratings are often modeled as being a combination of a latent, unobserved "true response" and a measurement error (Schmidt and Hunter, 1996). Applying this concept to peer review, Johnson (2008) proposed a model for ratings in which NIH reviewers' errors are defined by the extent to which their ratings tend to be higher or lower than other reviewers' on average. Johnson (2008) then analyzes how NIH's funding decisions would differ if they were to adjust for these measurement errors. Other approaches use multilevel regression modeling to account for differences in reviewers' average scores but do not explicitly characterize these differences as arising from measurement error (Erosheva et al., 2020; Jayasinghe et al., 2003).

In this vein, we can assess the refinement of estimated latent ("true") scores rather than the observed scores. We can even do so without explicitly estimating the latent scores: if we instead have an error distribution p_ϵ, we can solve the deconvolution $p_L \oplus p_\epsilon = p$ for p_L, the distribution of the latent scores (per Efron and Hastie (2016), this may be difficult). Entropic Refinement can then be computed for the distribution p_L. While the addition

[8]Rounding behavior will tend to decrease refinement but increase reliability (as rounding may turn disagreements into agreements, but not the other way around).

of independent noise increases entropy, Entropic Refinement is a difference of entropies, so latent score refinement may be higher or lower than that of the observed scores.

In our exposition of refinement and our application in Section 4, we use the observed scores and do not adjust for measurement error. Our approach aligns with the standard current practice of using unadjusted peer review scores.

With a small sample size of $N = 26$, we found moderate support for the hypothesis that AIBS merit scores—which are the only score used to make final funding decisions—are on average more refined than criterion scores. With a larger data set, more complex hypotheses about peer review informativeness could be tested with sufficient power. Consider the following hypothesis that could not be tested with currently available data:

Hypothesis 2. ***Reviewers who review applications they perceive as competitive display more refinement:*** When a reviewer believes an application's quality puts it near the funding cutoff, they may elect to expend the extra effort to distinguish that application from potential competitors by fine-tuning its scores.[9] This would manifest itself in higher entropy in rounding basins near a (perceived) funding cutoff.

Threshold-based incentives spur improved performance in other arenas,

[9]If this fine-tuning does not accurately reflect a reviewer's evaluation but rather stems from a desire to influence the proposal's funding outcome, it can be considered *gaming* (Coveney et al., 2017). Gaming is considered by some panelists to be unacceptable (Lamont, 2009).

e.g. ultramarathoning (Grant, 2016); we hypothesize that increased perceived likelihood of determining an application's funding similarly incentivizes reviewers to provide more refined ratings. According to an AIBS representative, there is no formal or informal "funding cutoff" known to reviewers, so data from a different funding institution and a survey of reviewers regarding their beliefs about a funding cutoff would be needed to test this hypothesis.

One limitation of Entropic Refinement is that it is specialized to decimal scales $\mathbb{S}$ such as the one used by AIBS. Our analysis reveals that such scales—in contrast to, for example, integer scales—provide raters with the ability to first conceptualize a round rating and then further refine. Nevertheless, extensions of refinement to other popular scale types are needed. The authors are currently investigating refinement for integer scales, such as the $\{1, \ldots 9\}$ scale used by NIH. With additional data, future analyses could assess whether refinement is greater for competitive-seeming applications, or could track reviewers over time to assess whether or not their scores' refinement increases as they gain experience. These types of studies will help illuminate the intricacies of ratings and human decision-making.

References

Athey, S. (2018). The Impact of Machine Learning on Economics. In *NBER Chapters*, pages 507–547. National Bureau of Economic Research, Inc.

Bailar, J. C. I. and Patterson, K. (1985). Journal Peer Review. *New England Journal of Medicine*.

Bar, H. Y. and Lillard, D. R. (2012). Accounting for heaping in retrospectively reported event data – a mixture-model approach. *Statistics in Medicine*, page 19.

Bopp, M. and Faeh, D. (2008). End-digits preference for self-reported height depends on language. *BMC Public Health*, 8(1):342.

Bornmann, L., Mutz, R., and Daniel, H.-D. (2010). A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. *PLoS ONE*, 5(12).

Coveney, J., Herbert, D. L., Hill, K., Mow, K. E., Graves, N., and Barnett, A. (2017). 'Are you siding with a personality or the grant proposal?': observations on how peer review panels function. *Research Integrity and Peer Review*, 2(1):19.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons. Google-Books-ID: VWq5GG6ycxMC.

Cunningham, W. H., Cunningham, I. C. M., and Green, R. T. (1977). The Ipsative Process to Reduce Response Set Bias. *The Public Opinion Quarterly*, 41(3):379–384. Publisher: [Oxford University Press, American Association for Public Opinion Research].

Curtice, J. (2009). Neither Representative nor Accountable: First-Past-the-Post in Britain. In *Duverger's Law of Plurality Voting: The Logic of Party Competition in Canada, India, the United Kingdom and the United States*, Studies in Public Choice, pages 27–45. Springer, New York, NY.

Drury, C. G. and Sinclair, M. A. (1983). Human and Machine Performance in an Inspection Task. *Human Factors*, 25(4):391–399.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge.

Erosheva, E., Walton, E. C., and Takeuchi, D. T. (2007). Self-Rated Health among Foreign- and US-Born Asian Americans: A Test of Comparability. *Medical care*, 45(1):80–87.

Erosheva, E. A., Grant, S., Chen, M.-C., Lindner, M. D., Nakamura, R. K., and Lee, C. J. (2020). NIH peer review: Criterion scores completely account for racial disparities in overall impact scores. *Science Advances*, 6(23):eaaz4868. Publisher: American Association for the Advancement of Science Section: Research Article.

Fang, F. C., Bowen, A., and Casadevall, A. (2016). NIH peer review percentile scores are poorly predictive of grant productivity. *eLife*, 5:e13323.

Feurer, I. D., Becker, G. J., Picus, D., Ramirez, E., Darcy, M. D., and Hicks,

M. E. (1994). Evaluating Peer Reviews: Pilot Testing of a Grading Instrument. *JAMA*, 272(2):98–100. Publisher: American Medical Association.

Gallo, S. (2021). Grant Peer Review Scoring Data with Criteria Scores. Publisher: figshare type: dataset.

Gallo, S. A., Sullivan, J. H., and Glisson, S. R. (2016). The Influence of Peer Reviewer Expertise on the Evaluation of Research Funding Applications. *PLOS ONE*, 11(10):e0165147.

Grada, C. (2006). Dublin Jewish Demography a Century Ago. *THE ECONOMIC AND SOCIAL REVIEW*, page 26.

Grant, D. (2016). The essential economics of threshold-based incentives: Theory, estimation, and evidence from the Western States 100. *Journal of Economic Behavior & Organization*, 130:180–197.

Grant, S., Meilă, M., Erosheva, E., and Lee, C. (2022). Refinement: Measuring informativeness of ratings in the absence of a gold standard. *British Journal of Mathematical and Statistical Psychology*, 75(3):593–615.

Green, J. G., Calhoun, F., Nierzwicki, L., Brackett, J., and Meier, P. (1989). Rating intervals: an experiment in peer review. *The FASEB Journal*, 3(8):1987–1992.

Hargens, L. and Herting, J. (1990). Neglected considerations in the analysis of agreement among journal referees. *Scientometrics*, 19(1-2):91–106.

Higginson, A. D. and Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biology*, 14(11):e2000995.

Jayasinghe, U. W., Marsh, H. W., and Bond, N. (2001). Peer Review in the Funding of Research in Higher Education: The Australian Experience. *Educational Evaluation and Policy Analysis*, 23(4):343–364.

Jayasinghe, U. W., Marsh, H. W., and Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 166(3):279–300.

Johnson, V. E. (2008). Statistical analysis of the National Institutes of Health peer review system. *Proceedings of the National Academy of Sciences of the United States of America*, 105(32):11076–11080.

Klesges, R. C., Debon, M., and Ray, J. W. (1995). Are self-reports of smoking rate biased? Evidence from the Second National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology*, 48(10):1225–1233.

Lamont, M. (2009). *How Professors Think*. Harvard University Press. Google-Books-ID: slK0xmSu33MC.

Langfeldt, L. (2001). The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. *Social Studies of Science*, 31(6):820–841.

Lauer, M. S., Danthi, N. S., Kaltman, J., and Wu, C. (2015). Predicting Productivity Returns on Investment. *Circulation Research*, 117(3):239–243.

Lauer, M. S. and Nakamura, R. (2015). Reviewing Peer Review at the NIH. *New England Journal of Medicine*, 373(20):1893–1895.

Lee, C. J. and Moher, D. (2017). Promote scientific integrity via journal peer review data. *Science*, 357(6348):256–257.

Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17.

Li, D. and Agha, L. (2015). Research funding. Big names or big ideas: do peer-review panels select the best science proposals? *Science (New York, N.Y.)*, 348(6233):434–438.

Lindner, M. D. and Nakamura, R. K. (2015). Examining the Predictive Validity of NIH Peer Review Scores. *PLoS ONE*, 10(6).

Lindner, M. D., Torralba, K. D., and Khan, N. A. (2018). Scientific productivity: An exploratory study of metrics and incentives. *PLOS ONE*, 13(4):e0195321.

Lynn, M., Flynn, S. M., and Helion, C. (2013). Do consumers prefer round prices? Evidence from pay-what-you-want decisions and self-pumped gasoline purchases. *Journal of Economic Psychology*, 36:96–102.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. University of Minnesota Press, Minneapolis, MN, US. Pages: x, 149.

Morera, O. F. and Dawes, R. M. (2006). Clinical and statistical prediction after 50 years: a dedication to Paul Meehl. *Journal of Behavioral Decision Making*, 19(5):409–412.

Nakamura, R. (2019). Testing of 2 Application Ranking Approaches at the National Institutes of Health Center for Scientific Review | Peer Review Congress.

Plug, C. (1977). Number Preferences in Ratio Estimation and Constant-Sum Scaling. *The American Journal of Psychology*, 90(4):699–704.

Schmidt, F. L. and Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1(2):199–223. Place: US Publisher: American Psychological Association.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., and Clark, L. (1991). RATING SCALES NUMERIC VALUES MAY CHANGE THE MEANING OF SCALE LABELS. *Public Opinion Quarterly*, 55(4):570–582.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K.,

and Wainwright, M. (2014). When is it Better to Compare than to Score? *arXiv:1406.6618 [cs, stat]*. arXiv: 1406.6618.

Shortliffe, E. H. and Sepúlveda, M. J. (2018). Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*, 320(21):2199–2200. Publisher: American Medical Association.

Simonsohn, U. (2013). Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone. *Psychological Science*, 24(10):1875–1888.

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9):160384.

Staff, M. S. L. (2019a). Ranked Choice Voting in Maine | Maine State Legislature.

Staff, N. (2012). Scoring System and Procedure.

Staff, N. (2019b). Enhancing Peer Review at NIH - Scoring and Review Changes.

Staff, N. (2019c). A Pilot Study of Half-Point Increments in Scoring | NIH Center for Scientific Review.

Stevenson, M. T. and Doleac, J. L. (2019). Algorithmic Risk Assessment in the Hands of Humans. page 72.

van Rooyen, S., Black, N., and Godlee, F. (1999). Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts. *Journal of Clinical Epidemiology*, 52(7):625–629.

Wang, J., Veugelers, R., and Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436.

Wilcoxon, F. (1946). Individual Comparisons of Grouped Data by Ranking Methods. *Journal of Economic Entomology*, 39(2):269–270. Publisher: Oxford Academic.

A Proofs

This section provides proofs of propositions from the paper.

A.1 Proposition 1: Refinement Decomposition

We require the following Lemma, which is a generalization of (Cover and Thomas, 2012) Chapter 2, Exercise 19:

Lemma 1. *For a finite mixture P of m distributions with mutually disjoint support, $P = \sum_{i=1}^{m} \lambda_i P_i$ with $\sum_i \lambda_i = 1$ and $\lambda_i \geq 0$ for all i,*

$$H(P) = -\sum_i \lambda_i \log \lambda_i + \sum_i \lambda_i H(P_i) \tag{6}$$

Proof: We have

$$\begin{aligned} H(P) &= -\sum_{i=1}^{m} \sum_{k \in supp(P_i)} \lambda_i P_i(k) \log(\lambda_i P_i(k)) \\ &= -\sum_i \lambda_i \log(\lambda_i) + \sum_i \lambda_i H(P_i). \end{aligned}$$

The second equality follows from the log of a product equaling the sum of the logs and rearrangement of terms. We can now derive the refinement decomposition.

Proof of Proposition 1: Taking $0 \log(0) = 0$ and δ_s the degenerate distribution with $\mathrm{P}(s) = 1$,

$$\begin{aligned} H(R_t(\mathbf{Y})) &= H(\sum_{s \in \mathbb{S}_t} p(B_t(s)) \delta_s) \\ &= -\sum_{s \in \mathbb{S}_t} p(B_t(s)) \log(p(B_t(s))) \end{aligned} \tag{7}$$

by Lemma 1 since B_t generates a partition and hence empirical distributions with disjoint support, and the entropy of degenerate distributions δ_s is zero. Applying the same lemma to the unrounded $\mathbf{Y}$ yields, in a similar fashion,

$$H(Y) = -\sum_{s \in \mathbb{S}_t} p(B_t(s)) \log(p(B_t(s))) + \sum_{s \in \mathbb{S}_t} p(B_t(s)) H(\mathbf{Y}_s).$$

Combining with (7) finishes the proof.

A.2 Proposition 2: Extrema and Range

Proposition 2 follows from Proposition 1. It is clear from Proposition 1 that Entropic Refinement achieves its minimum of 0 when there is only one unique observed score value in any basin of attraction, so that $H(\mathbf{Y}_s) = 0$ for all s for which $p(B(s)) > 0$. With a bit more effort we can construct the maximum. Because within-basin entropy only depends on the scores in a single basin, the $H(\mathbf{Y}_s)$ terms in (4) can be independently maximized. The uniform distribution over k elements yields maximum entropy $\log(k)$ for discrete distributions over k outcomes, so we must have within-basin uniformity. Finally, (4) is a weighted average, so the maximum occurs when only maximum-sized basins have nonzero weight. Thus the maximizing empirical distributions put probability mass only on the maximum-sized basins. Maximum Entropic Refinement therefore occurs when the empirical score distribution is uniform on maximum-size basins and there are no observed scores in other basins.

A.3 Proposition 3: Joint vs. Average Entropic Refinement

This section proves Proposition 3 for (WLOG) $t = 1$. We first prove that $r_E^{avg} \leq r_E^{joint}$. Recall that C is the number of criterion scores, so that

$$H(\mathbf{X}) \geq \max_{i \in [C]} H\left(\mathbf{X}^i\right) \geq \frac{1}{C} \sum_{i=1}^{C} H\left(\mathbf{X}^i\right) \tag{8}$$

where the first inequality is equality only if the criterion scores are deterministically related.

Now let $\mathbf{X}_s \equiv \left\{x \in \mathbf{X} \cap B_1^C(s)\right\}$ where $B_1^C(s)$ is the C-dimensional basin of attraction to $s \in \mathbb{S}_1^C$. Let i index criteria, so that $\mathbf{X}_s^i$ is the set of scores on criterion i in the rounding basin of s. We then write the joint entropy in terms of the decomposition (4) and substitute (8):

$$
\begin{aligned}
r_E^{joint}(\mathbf{X}) &= \sum_{s \in \mathbb{S}_1^C} p(\mathbf{X}_s) H(\mathbf{X}_s) && (9)\\
&\geq \sum_{s \in \mathbb{S}_1^C} p(\mathbf{X}_s) \frac{1}{C} \sum_{i \in [C]} H\left(\mathbf{X}_s^i\right) && (10)\\
&= \frac{1}{C} \sum_{i \in [C]} \sum_{s \in \mathbb{S}_1^C} p(\mathbf{X}_s^i) H\left(\mathbf{X}_s^i\right) && (11)\\
&= r_E^{avg}(\mathbf{X}) && (12)
\end{aligned}
$$

(the third line follows simply from distributing $p(\mathbf{X}_s)$ inside the sum over i and rearranging sums).

For the second inequality, $Cr_E^{avg} \geq r_E^{joint}$, denote p^i the empirical probability distribution associated to $\mathbf{X}^i$, for $i \in [C]$. Assume initially that $C = 2$.

Consider first the case $\mathbf{X}^1 \perp\!\!\!\perp \mathbf{X}^2$, that is, $p = p^1 p^2$. In this case, we have that $H(\mathbf{X}) = H\left(\mathbf{X}^1\right) + H\left(\mathbf{X}^2\right)$. Moreover, because in this case $R\left(\mathbf{X}^1\right) \perp\!\!\!\perp R\left(\mathbf{X}^2\right)$, we also have $H(R(\mathbf{X})) = H\left(R\left(\mathbf{X}^1\right)\right) + H\left(R\left(\mathbf{X}^2\right)\right)$. It follows then, immediately, that $r_E^{joint} = \frac{1}{2}\left(r_E\left(\mathbf{X}^1\right) + r_E\left(\mathbf{X}^2\right)\right) = r_E^{avg}$.

Let us now consider the general case, for which

$$H(\mathbf{X}) = H\left(\mathbf{X}^1\right) + H\left(\mathbf{X}^2\right) - I\left(\mathbf{X}^1, \mathbf{X}^2\right),$$

where the last term denotes the *mutual information* (Cover and Thomas, 2012) between the two criterion scores. Similarly,

$$H(R(\mathbf{X})) = H\left(R\left(\mathbf{X}^1\right)\right) + H\left(R\left(\mathbf{X}^2\right)\right) - I\left(R\left(\mathbf{X}^1\right), R\left(\mathbf{X}^2\right)\right),$$

and therefore

$$Cr_E^{joint} = r_E\left(\mathbf{X}^1\right) + r_E\left(\mathbf{X}^2\right) - \left(I\left(\mathbf{X}^1, \mathbf{X}^2\right) - I\left(R\left(\mathbf{X}^1\right), R\left(\mathbf{X}^2\right)\right)\right).$$

We show now that the last term, $I\left(\mathbf{X}^1, \mathbf{X}^2\right) - I\left(R\left(\mathbf{X}^1\right), R\left(\mathbf{X}^2\right)\right)$, is non-negative. For this we use the data processing inequality (Cover and Thomas, 2012), which states that if random variables X, Y, Z form a Markov chain, in other words if $X \perp\!\!\!\perp Z|Y$, then $I(X,Y) \geq I(X,Z)$. We have that $\mathbf{X}^1 \perp\!\!\!\perp R(\mathbf{X}^2)|\mathbf{X}^2$, from which we derive that $I\left(\mathbf{X}^1, \mathbf{X}^2\right) \geq I\left(\mathbf{X}^1, R\left(\mathbf{X}^2\right)\right)$. Moreover, $R\left(\mathbf{X}^2\right) \perp\!\!\!\perp R\left(\mathbf{X}^1\right)|\mathbf{X}^1$, from which we have that $I\left(\mathbf{X}^1, R\left(\mathbf{X}^2\right)\right) \geq I\left(R\left(\mathbf{X}^1\right), R\left(\mathbf{X}^2\right)\right)$, which concludes the proof. The proof for $C > 2$ follows by induction.