The Magnitude of Truth: On Using Magnitude Estimation for Truthfulness Assessment

Michael Soprano
University of Udine
Udine, Italy
michael.soprano@uniud.it

Denis Eduard Tapu
University of Udine
Udine, Italy
tapu.deniseduard@spes.uniud.it

David La Barbera
University of Udine
Udine, Italy
david.labarbera@uniud.it

Kevin Roitero
University of Udine
Udine, Italy
kevin.roitero@uniud.it

Stefano Mizzaro
University of Udine
Udine, Italy
stefano.mizzaro@uniud.it

Abstract

Assessing the truthfulness of information is a critical task in fact-checking, and is typically performed using binary or coarse ordinal scales (2–6 levels), though fine-grained scales (e.g., 100 levels) have also been explored. Magnitude Estimation (ME) takes this approach further by allowing assessors to assign any value in the range $(0, +\infty)$. However, it introduces challenges, including the need for aggregation of assessments from individuals with different interpretations of the scale. Despite these, its successful applications in other domains suggest its potential suitability for truthfulness assessment. We conduct a crowdsourcing study by collecting assessments on claims sourced from the PolitiFact fact-checking organization using ME. To the best of our knowledge, this is the first systematic investigation of ME in the context of truthfulness assessment. Our results show that while aggregation methods significantly impact assessment quality, optimal aggregation strategies yield accuracy and reliability comparable to traditional scales. More importantly, ME allows capturing subtle differences in truthfulness, offering richer insights than conventional coarse-grained scales.

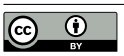
CCS Concepts

• **Information systems** → **Crowdsourcing**; **Clustering and classification**; • **General and reference** → **Metrics**.

Keywords

Misinformation, Crowdsourcing, Fact-checking

ACM Reference Format:
Michael Soprano, Denis Eduard Tapu, David La Barbera, Kevin Roitero, and Stefano Mizzaro. 2025. The Magnitude of Truth: On Using Magnitude Estimation for Truthfulness Assessment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3730091



SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730091

1 Introduction

Misinformation is a widespread issue that continues to intensify, with its spread further exacerbated by social media [7]. One countermeasure is fact-checking, i.e., the verification of facts [16]. Traditionally, it has been performed by expert journalists who are part of well-known organizations [54, 64]. The fact-checking process involves several activities, a crucial one being the assessment of the truthfulness of the information item examined [36]. Recent research has attempted to scale up fact-checking by outsourcing truthfulness assessments to non-experts through crowdsourcing approaches [1, 2, 10, 24, 25, 33, 40, 44, 46, 48, 49, 55].

A specific but important issue, on which we focus in this work, is the scale adopted to express and represent the truthfulness of an information item. At first sight, truthfulness assessment might simply seem like a binary classification task, with each information item being labeled as true or false; however, the notion of "truth" in fact-checking is perhaps not so clear and shows several nuances that cannot be captured by a binary scale. Indeed, whether the task is performed by experts or by means of crowdsourcing, several different scales have been adopted. These range from coarse-grained (e.g., 2-3 labels) to fine-grained (e.g., up to 100 levels). Fact-checking organizations adopt varied approaches: RMIT ABC Fact Check[1] uses a three-level scale, while PolitiFact[2] employs six levels.

Researchers have adopted an even wider variety of approaches: scales with 2, 3, 5, 6, and even 101 levels have been used [6, 10, 24–26, 44, 46–49, 55]. Notice that the change from a two-level scale to a scale with more levels implies a fundamentally different problem, as it should be framed as an ordinal classification problem, for which universally accepted metrics are not yet established [3, 4, 46]. Additionally, multidimensional scales have been tried [55], which adds another level of complexity.

In the literature, there is also a peculiar scale called Magnitude Estimation (ME), originally proposed by Stevens [57, 58]. Such a scale involves an assessor numerically estimating the perceived strength (magnitude) of a stimulus along a dimension of interest. The assessor allocates arbitrary numbers, typically in the $(0, +\infty)$ range, creating a ratio scale of perception. For instance, if a magnitude of 1 is assigned to one stimulus and 0.5 to another, the second is perceived as half as strong. A magnitude of 10 for a third stimulus

[1] https://www.abc.net.au/news/factcheck
[2] https://politifact.com

indicates it is perceived as ten times stronger than the first. The assessor can always find a larger or smaller number as needed (since zero cannot be used). The scale can be normalized by dividing all values by the largest perceived value [22].

Despite several challenges (discussed in the next section), ME has had many successful applications, such as collecting relevance assessments via crowdsourcing [28, 29, 51, 61, 65], a task that shares similarities with using crowdsourcing for truthfulness assessments. This situation motivates the study we present here. To the best of our knowledge, ME has not been systematically applied in this setting, with only a preliminary study by Roitero et al. [44] addressing it. Their work, a workshop paper from 2018, provides an initial exploration of the topic but is now outdated and has several limitations. Specifically, they compared a 100-level scale with an ME-based scale, collecting 800 assessments for each, but relied on MTurk, a platform facing concerns over declining worker quality [20]. We improve upon their experimental setup by (i) collecting twice the number of statements and crowd assessments, (ii) recruiting higher-quality workers from Prolific, a platform shown to yield superior assessments [18, 21, 32] and widely used in truthfulness research [24], and (iii) conducting a deeper and broader analysis of the collected data. These improvements enable us to provide more comprehensive and reliable findings. Specifically, we address the following Research Questions:

RQ1 Is ME a feasible scale for collecting truthfulness assessments in a crowdsourcing setting? Can assessments made using ME align with expert fact-checkers' ground truth, as measured by a traditional coarse-grained scale?

RQ2 How does ME compare to traditional coarse-grained scales? How do ME-based assessments compare to those collected using a traditional coarse-grained scale under the same conditions?

RQ3 Does ME provide more nuanced information than traditional coarse-grained scales? Does it capture subtleties in truthfulness that traditional scales may overlook?

To address these RQs, we conduct a crowdsourcing study to collect truthfulness assessments using ME, comparing them to those obtained with a six-level scale under the very same experimental conditions by La Barbera et al. [24]. All data, code, task configuration, and supplementary materials are publicly available.[3]

Our contributions are threefold. First, we provide evidence supporting the feasibility of using ME for collecting truthfulness assessments, demonstrating through a large-scale crowdsourcing study its practicality and robustness in this context, also when compared to more traditional scales. Second, we show how different aggregation strategies impact the quality and reliability of truthfulness assessments, and identify the optimal techniques for leveraging ME effectively. Third, we highlight how ME allows for a deeper understanding of truthfulness perception by revealing patterns and trends that are less apparent with traditional coarse-grained scales.

The remainder of this work is structured as follows: Section 2 addresses previous ME applications, focusing on crowdsourcing and truthfulness assessment. Section 3 outlines the methodology used to address the research questions. Section 4 presents and discusses the results, and Section 5 concludes the paper.

[3] https://doi.org/10.17605/OSF.IO/YUX42

2 Related Work

We highlight some potential issues when using ME in Section 2.1. We then explore previous approaches that use ME in Section 2.2, focusing on crowdsourcing. In Section 2.3, we examine studies that collect truthfulness assessments through crowdsourcing.

2.1 Challenges in Using Magnitude Estimation

Using Magnitude Estimation (ME) for general tasks presents challenges. Assessors may find the method difficult to grasp, as it seems unfamiliar and requires advanced mathematical understanding, without offering clear advantages over traditional coarse-grained scales [14]. If assessors know the magnitude of the first stimulus, they may anchor their responses to this reference, introducing bias [39]. This skew results in over-represented values [37]. Stevens [58] points out that assessors interpret the scale differently, which complicates comparisons and interpretations. Small changes in stimulus intensity can also result in significant perceived differences, leading to nonlinear response patterns.

When applying ME to truthfulness assessment, these challenges become more pronounced. The unbounded scale lacks meaningfulness because truthfulness does not inherently imply stimulus intensity. Its fine-grained nature can confuse assessors, making it difficult to define small deviations from the truth and potentially encouraging over-interpretation. Subjectivity, influenced by prior knowledge or biases, leads to greater variability compared to other domains. Ambiguity in items, opinions, or language further complicates assigning precise truthfulness levels. Systematic biases, as Soprano et al. [54] describes, exacerbate errors when assessors favor certain sources or topics. Despite these obstacles, successful uses of ME highlight its potential for truthfulness assessment, demonstrating the importance of exploring its complexities.

2.2 Applications of Magnitude Estimation

Despite the potential issues, researchers have applied ME in various fields. Kemp [19] studied its application in public administration research to examine the perceived utility of government-supplied services. Engen and McBurney [11] used it in olfactology to assess the pleasantness of odors, while Rutschmann [50] investigated the perception of anxiety and related phenomena. Another field of application is linguistics, where it is used to collect acceptability assessments [5]. Featherston [12] focused on estimating how sentences adhered to the grammatical rules of German, while Sorace and Keller [56] examined the degrees of acceptability of sentences.

There are also applications of ME in psychophysics; Marks [31] and Fucci et al. [13] used it to obtain sensory matches for the loudness of stimuli, while McDaniel et al. [34] focused specifically on sonic booms. Another field of application is human-computer interaction. Shinobu et al. [52] used ME to estimate the overall usability of a target product in a quantitative way, while McGee [35] focused on software interface usability, and Rich and McGee [43] addressed user expectations in usability tests.

When applying ME in crowdsourcing, researchers have used these methods to gather document relevance assessments for test collection-based evaluations of information retrieval systems. Maddalena et al. [28] conducted a preliminary study, collecting 960

crowdsourced assessments using ME to evaluate document relevance. Building on this work, Turpin et al. [61] and Maddalena et al. [29] expanded the approach, gathering over 50,000 ME-based relevance assessments. In their updated settings, workers first completed a warm-up task, estimating the magnitudes of three lines of different lengths shown sequentially, before assessing the relevance of eight documents. Additionally, Yang et al. [65] compared relevance assessments gathered using three different crowdsourcing techniques with those obtained through ME in prior studies.

Roitero et al. [45] explored the differences and effects of using four scales for crowdsourcing relevance assessments: a two-level scale, a four-level scale, a 101-level scale, and an ME-based scale. They demonstrated that ME is less suitable in applications where high agreement among workers is critical, as it tends to produce lower agreement with expert-provided assessments compared to more coarse-grained scales.

All studies indicate that ME-based assessments are reliable and align consistently with categorical expert-provided assessments, showing no notable drawbacks related to document ordering. Furthermore, Maddalena et al. [29] demonstrated through a failure analysis that discrepancies between expert and non-expert assessments are not due to ME but are influenced by other factors. Notably, only a limited number of crowdsourced assessments are required to achieve stable results.

2.3 Crowdsourcing Truthfulness Assessments

Several researchers have recruited crowds of non-experts from crowdsourcing platforms to assess the truthfulness of information items, typically using a six-level coarse-grained assessment scale.

La Barbera et al. [25] collected 1,200 truthfulness assessments for 120 information items made by public figures to examine the impact of judgment scales and assessors' backgrounds. Similarly, Roitero et al. [46] used a comparable setup for 180 items. Roitero et al. [49] gathered truthfulness assessments for COVID-19-related items and conducted a longitudinal study [48] on the effects of information recency and time. Draws et al. [10] used a similar setup to explore cognitive biases, while Soprano et al. [55] assessed truthfulness across seven dimensions. An exception is Roitero et al. [46], who used three scales and discussed their properties in detail. Overall, these studies suggest that the six-level scale allows crowd workers to provide truthfulness assessments that align with expert assessments under certain conditions.

3 Methodology

We describe the dataset (Section 3.1), the design of the crowdsourcing task (Section 3.2), and the assessment collections (Section 3.3). We then explain the scale transformations (Section 3.4), groupings (Section 3.5), aggregation functions (Section 3.6), and the measures used to evaluate assessment effectiveness (Section 3.7).

3.1 Dataset

We use information items from the PolitiFact organization for our crowdsourcing task. Since 2007, PolitiFact has fact-checked claims made by U.S. politicians, political organizations, public figures, and on social media. With over 24,000 fact-checks available, the website is regularly updated. Soprano et al. [54] provided a detailed description of the organization's process. Truthfulness assessments use a six-level ordinal scale (S_6): Pants On Fire, False, Mostly False, Half True, Mostly True, and True.

Table 1: Sample of two PolitiFact information items.

Speaker	Date	Statement	Ground Truth
Ted Nugent	June 14, 2022	*Three mass shootings were meant to distract from Hillary Clinton controversies.*	Pants On Fire
Levar Stoney	June 16, 2022	*In Virginia, Black people are eight times (8X) more likely than white people to die of gun homicide.*	True

For direct comparison, we use the same information items as La Barbera et al. [24]. Our dataset includes 120 items, evenly distributed across the six ground truth levels, with 20 items per level. Of these, 65 items were issued by speakers or organizations affiliated with the Republican Party, and the remaining 55 were from the Democratic Party. Table 1 presents two sample items from the ends of the scale.

3.2 Crowdsourcing Task

The task workflow is as follows: each crowd worker is first presented with instructions that describe the task and provide a definition of ME, including an example. They are then asked to complete a demographic questionnaire consisting of five questions to gather background information. As part of the training, workers are asked to estimate the length of three lines using ME, with each line presented one at a time in random order. This training, inspired by Maddalena et al. [29], familiarizes the workers with the scale.

Once trained, workers assess eight information items, one at a time and in random order, to minimize biases and learning effects. The assessment interface, based on La Barbera et al. [24], follows three distinct sequential steps. First, the workers read the statement. Then, they search for evidence using a customized search engine. After selecting one of the retrieved results, the workers provide the truthfulness assessment using ME by entering a number in the range $(0, +\infty)$. Once the task is completed, the workers fill out a questionnaire with four questions aimed at assessing their perceived usability of ME.

Overall, we used a total of 120 information items (Section 3.1), with each statement evaluated by 10 distinct crowd workers. Each worker assesses the truthfulness of 8 items, two of which are designed to be clearly true and false, serving as gold questions. Additionally, we require each worker to spend at least two seconds on the training elements and three seconds on each statement and questionnaire to ensure data quality.

We published our task on the Prolific platform, as done by La Barbera et al. [24], which Douglas et al. [9] have shown to yield higher-quality data compared to other crowdsourcing platforms like Amazon Mechanical Turk. The task ran for three weeks, with 200 work units published and 2,200 assessments collected. We set a completion time of 20 minutes with a payment of £2.50 (approximately $3.10), corresponding to an hourly rate of £7.50 (about $9.40), based on an initial pilot test. In practice, the average task completion time was 19 minutes (σ = 11 mins, median = 15.5 mins),

resulting in a median hourly reward of approximately £10 (about $12.50).

We designed and conducted the crowdsourcing task using the Crowd_Frame framework, which facilitates the deployment of crowdsourcing experiments [53].

3.3 Assessment Collections

We consider two sets of crowdsourced assessments: 1,200 collected using ME, referred to as $\mathsf{ME}_{\mathsf{Crowd}}$, and 1,200 collected by La Barbera et al. [24] using the same experimental design but with a six-level scale (S_6), referred to as $\mathsf{S6}_{\mathsf{Crowd}}$.

After filtering out training assessments and those provided for gold questions from both sets, we compare them with the 120 ground truth assessments published by PolitiFact, denoted as GT. Both $\mathsf{S6}_{\mathsf{Crowd}}$ and GT use the labels Pants On Fire, False, Mostly False, Half True, Mostly True, and True.

3.4 Normalization

We normalize and group the assessments to compare the two scales, ME and S_6, as outlined by Roitero et al. [45, 46]. Normalization standardizes the scales used by different individuals [37]. While various strategies exist, geometric averaging is often recommended, as noted by Gescheider [15], Moskowitz [37], and McGee [35]. This method has also been used for relevance assessments by Maddalena et al. [29]. The assumption is that ME assessments are log-normally distributed. However, to align the assessments in $\mathsf{ME}_{\mathsf{Crowd}}$ with the S_6 range of $[0, 5]$, we adopt a simpler normalization approach that preserves the ratio information.

We then consider the lowest and highest assessments provided by the worker for the information items in each work unit. These assessments define the virtual lower and upper bounds of the relative scale [45]. Next, we apply a min-max transformation to adjust the assessments in $\mathsf{ME}_{\mathsf{Crowd}}$ to the S_6 range. Let A_w represent the set of assessments provided by worker w for the information items in their assigned work unit. For an individual assessment $\alpha \in A_w$, the normalized assessment is defined as $\alpha_{\text{norm}} = \frac{\alpha - \min(A_w)}{\max(A_w) - \min(A_w)} \cdot 5$, where $\min(A_w)$ and $\max(A_w)$ are the minimum and maximum values in A_w. This normalization scales the assessments to a range of 0 to 5.

3.5 Groupings

In certain analyses, it is beneficial to look at the assessments by grouping the original ground truth scale levels into fewer categories, as described by Soprano et al. [55] and Roitero et al. [48].

To further compare S_6 and ME, we thus perform two groupings based on the ground truth levels (GT). The original ground truth with six levels is denoted as GT_6. In the first grouping, we combine the first three levels of the ground truth (Pants On Fire, False, Mostly False) into a single level, while the last three levels (Half True, Mostly True, True) form another. We assume that the labels from both initial scales are evenly distributed, making the midpoint a suitable dividing threshold. This resulting group is denoted as GT_2. In the second grouping, we combine Pants On Fire with False, Mostly False with Half True, and Mostly True with True, referring to this group as GT_3.

More specifically, the original ground truth with six levels, denoted as GT_6, comprises six categorical ordinal truthfulness labels that map to the numeric values $\{0, 1, 2, 3, 4, 5\}$ in ascending order of truthfulness. In GT_2, these labels are grouped into $\{0, 1, 2\}$, while in GT_3, they are grouped into $\{0, 1\}$.

3.6 Aggregation Functions

We aggregate individual ME assessments for each information item into a single value, a common practice in crowdsourcing. Specifically, we combine the truthfulness assessments from multiple workers into a single representative score. This aggregation allows us to summarize the collective assessment, providing a more reliable and robust measure of the information item's truthfulness. By applying different aggregation techniques, we ensure the resulting score reflects a balanced consensus while accounting for individual variability. As done by Maddalena et al. [29], we compare various aggregation functions to identify the most effective: the geometric mean (gmean), the arithmetic mean (mean), the weighted mean (wmean, with specific weights discussed in Section 3.7), and the median.

The arithmetic mean has shown the best results in studies on ME-based relevance assessments [45] and on truthfulness assessments using a six-level scale [46]. In the context of truthfulness assessment, La Barbera et al. [24] applied all of these aggregation functions (but not on ME data). As discussed later, while the gmean is theoretically appropriate for ME scales [37], it, along with the median, performs worse than other aggregation functions due to its inability to effectively handle normalized truthfulness values of 0. In contrast, the wmean emerges as the most effective aggregation function for our data, achieving the highest scores across all the measures we consider, except for one (Section 4.2). Although no single aggregation rule is universally optimal, the weighted mean is a suitable choice for real-world assessments because it reflects internal agreement and produces robust, stable aggregates.

3.7 Measures

To evaluate the effectiveness of the truthfulness assessment activity, we employ the same set of measures defined by La Barbera et al. [24]. For a detailed discussion, see their Section 4.4. In brief, we address an ordinal classification problem that differs from simple classification, since misclassifications vary in severity, and regression, since categories are not equidistant. For example, assigning a score of 2 when the ground truth is 1 is less severe than assigning a score of 5. As Amigó et al. [3, 4] highlight, no standard measures exist for such a case; therefore, we rely on a combination of three types of measures: accuracy, agreement, and error values. We compute each measure in two versions:

- *Individual*: We calculate the measure for the 6 non-gold items that each worker assesses in their work unit.
- *Aggregated*: We determine the measure (excluding Internal agreement) from the aggregated assessments of 120 non-gold items, producing a single score for each measure.

To evaluate accuracy, we compare the crowdsourced assessments with the experts' ground truth. Since we group the ground truth into fewer levels, we calculate accuracy for the original six levels

(GT_6), the grouped three levels (GT_3), and the grouped two levels (GT_2). We denote the resulting accuracy scores as $Accuracy_6$, $Accuracy_3$, and $Accuracy_2$, respectively, with Accuracy representing the general case. These accuracy scores are computed for each set of crowdsourced assessments we analyze, ME_{Crowd} and $S6_{Crowd}$. By reporting effectiveness on a binary categorization problem, we aim to facilitate comparisons with other state-of-the-art approaches [46].

We also compute external agreement (External) to measure the alignment between workers' assessments and expert labels. The assumption is that if the crowd's assessments align well with expert labels, the effectiveness of the assessments is high. Additionally, we consider internal agreement (Internal) among workers. This score is computed exclusively from the assessments provided by workers, without considering those of experts. Here, the assumption is that higher internal agreement reflects higher effectiveness. For both internal and external agreement, we use the α reliability coefficient by Krippendorff [23]. The Internal agreement scores also serve as weights in the wmean aggregation function (see Section 3.6).

We also compute pairwise agreement (Pairwise), as defined by Maddalena et al. [30]. This measure represents the number of cases where workers agree, divided by the total number of observations. Intuitively, it quantifies the fraction of pairs in agreement between a "ground truth" scale and a "crowd" scale [46]. Specifically, a pair of truthfulness assessments is considered in agreement if one assessment is lower than the other and its ground truth is also lower. Similar to the accuracy calculation, we compute pairwise agreement for each grouping of the ground truth, denoting each measure as $Pairwise_6$, $Pairwise_3$, and $Pairwise_2$ (or Pairwise in general).

Finally, we compute the mean squared error (MSE) to quantify the variance between workers' assessments and the ground truth, providing a numerical measure of accuracy. Additionally, we consider the mean absolute error (MAE) and square error (SE). While MAE is robust, easy to interpret, and preserves the unit of measurement, MSE is sensitive to outliers, penalizing larger errors more heavily and resulting in values that are less interpretable. This sensitivity allows MSE to potentially capture different signals than MAE.

4 Results

We present descriptive statistics about the crowd workers and the collected data (Section 4.1). We then study the alignment of ME-based assessments with expert ones (RQ1), compare ME with traditional coarse-grained scales (RQ2), and investigate whether ME reveals additional insights missed by traditional scales (RQ3).

4.1 Descriptive Statistics

We begin by presenting the demographic characteristics of the workers who completed our task, along with data on the number who completed, failed, or abandoned it.

From the questionnaire responses, we derived the following demographic statistics: the most frequent age range is 36–50 (37.5%, 75/200). Regarding education, 39.5% (79/200) hold a four-year college degree or higher. In terms of family income before taxes last year, 21.5% (43/200) earned between $50,000 and $75,000. On political views, 34% (68/200) identify as liberal, while 48.5% (97/200) identify as Democrats. The Prolific platform provides additional demographic data with worker consent. The majority of workers identify as male (53%, 106/200) and as white (53%, 106/200). All workers are based in the USA, with most born there (79.5%, 159/200), and all hold full citizenship. English is the first language for 89.5% (179/200) of workers. Additionally, most workers are not students (67%, 134/200), and roughly half are employed full-time (50.5%, 101/200). Overall, our sample is well-balanced demographically, consistent with previous studies summarized in Section 2.3, including those by Roitero et al. [46] and La Barbera et al. [24].

Out of 332 workers who participated in the task, 200 (60.24%) successfully completed the published work units. Of the remaining 132 workers (39.76%), 103 (78.03%) abandoned the task during their first or subsequent attempts, while 29 (21.97%) failed without retrying. A small number of workers (9, 6.82%) who completed the work units also failed the training task, with 2 failing twice and the others once. The overall abandonment rate, defined by Han et al. [17], aligns with those reported in previous studies [10, 24, 25, 46, 48, 49, 55], as do the demographic characteristics.

4.2 RQ1: Alignment of ME with Ground Truth

We begin by investigating the distributions of the 1,200 assessments collected using ME, denoted as ME_{Crowd}. The frequency distribution of individual assessments is shown in Figure 1 (first two plots). ME scores generally follow a log-normal distribution [37], as demonstrated by Maddalena et al. [29] and Roitero et al. [44]. However, the ME_{Crowd} assessments deviate from this pattern, as indicated by the D'Agostino-Pearson normality test ($p > 0.05$) [8, 62, 63].

When analyzing the distribution and usage patterns of ME_{Crowd}, we observe that 15.25% of the individual truthfulness assessments fall within the decimal range $(0, 1)$, while the majority (84.06%) fall within the range $[1, 100]$. Among workers, 39% use at least one value in the $(0, 1)$ range, while 59.5% provide all their assessments using only values in the $[1, 100]$ range. Another characteristic of ME is the rounding bias, where assessors tend to favor round numbers when making estimates [37]. For ME_{Crowd}, 33.06% of the assessments are multiples of 5. This indicates a tendency toward round numbers, consistent with the findings of Roitero et al. [45].

Before commenting on the last two plots of Figure 1, we first discuss the aggregation of individual assessments. Table 2 compares the performance of different aggregation functions with respect to each measure, excluding Internal. As shown, and as anticipated in Section 3.6, wmean achieves the best scores for all measures except External, where median performs better. Consequently, we will primarily report results using wmean as the aggregation function. The Internal metric is not aggregated; instead, it is used to compute the weights for wmean by normalizing scores into the $[0, 1]$ range. This gives more weight to workers who agree more with others when assessing the same information items, a common technique in crowdsourcing [27].

After discussing the aggregation methods, we return to the last two plots in Figure 1, which show the distributions of normalized individual truthfulness assessments and aggregated assessments. As shown, aggregating the assessments results in a distribution that more closely resembles a bell-shaped curve, and unlike the

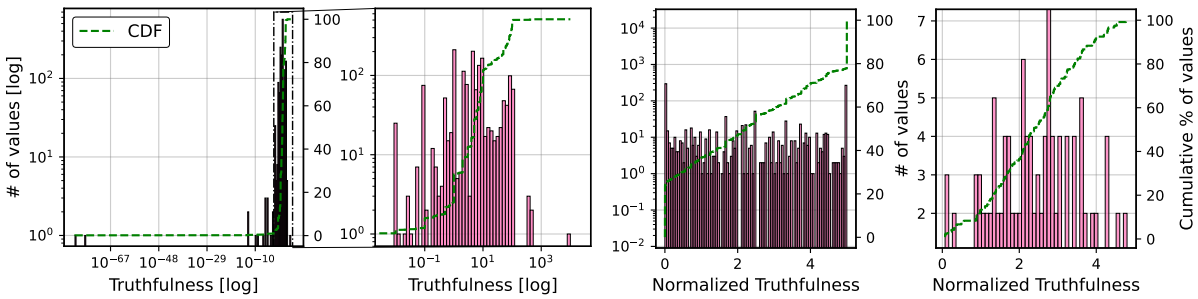


Figure 1: Distributions of truthfulness assessments (ME_{Crowd}). First plot: individual assessments. Second plot: 99% of the individual assessments. Third plot: individual normalized assessments. Fourth plot: aggregated normalized assessments.

Table 2: Comparison of aggregation functions across effectiveness measures.

	$Accuracy_2$	$Accuracy_3$	$Accuracy_6$	External	$Pairwise_2$	$Pairwise_3$	$Pairwise_6$	MAE	MSE
wmean	**0.80**	**0.60**	**0.32**	0.63	**0.89**	**0.83**	**0.79**	**0.97**	**1.59**
median	0.73	0.59	0.28	**0.67**	0.87	0.82	0.78	1.02	1.73
mean	0.75	0.57	0.29	0.53	0.86	0.82	0.77	1.07	1.78
gmean	0.57	0.41	0.21	-0.15	0.26	0.25	0.24	2.10	7.06

individual assessments, it is normally distributed ($p < 0.05$). This supports common findings highlighted in previous studies [44, 45].

We now discuss the alignment of the aggregated ME_{Crowd} assessments with the ground truth (GT) and compare them with $S6_{Crowd}$. Figure 2 shows how each statement corresponds with the ground truth, along with a breakdown for each ground truth value. The x-axis represents the ground truth values, and the y-axis shows the aggregated truthfulness values. Each circle represents a statement, while each triangle indicates the mean value. The blue color represents ME, and the gray color represents S_6. The left plot corresponds to the GT_6 grouping, the center plot to GT_3, and the right plot to GT_2.

We begin with ME (blue). The aggregated assessments align with the ground truth: the boxplots show median values consistent with expert assessments, rising with the truthfulness level. For GT_2, grouping the ground truth into two levels, the accuracy is 0.80, allowing comparison with automatic truthfulness systems. Thorne et al. [59] reported 0.50, Popat et al. [41] reported 0.56, Nakashole and Mitchell [38] achieved 0.78, and Potthast et al. [42] reported 0.75. Our accuracy is comparable to or better than these systems, which show significant variation across datasets. The result obtained using ME is also comparable to the accuracy of S_6 (gray), which is 0.82.

Focusing on the ME–S_6 comparison, the trends show that both achieve similar effectiveness. Starting with the GT_6 grouping, ME performs better for Pants On Fire, False, and Mostly False, with lower aggregated truthfulness values closer to the ground truth false level. In contrast, S_6 outperforms ME for Half True, Mostly True, and True, showing narrower boxplots and more distant medians. Similar trends are observed for GT_3 and GT_2, with statistically significant differences. For GT_6, the statistically significant levels are False and Mostly False. For GT_3, these levels are Pants On Fire and False, while for GT_2, they are Pants On Fire, False, and Mostly False. Given the similar effectiveness of the two scales, we confirm the reliability of the collected assessments.

4.3 RQ2: Comparing ME with Traditional Scales

After studying the alignment of ME_{Crowd} with the ground truth (GT), we compare it with $S6_{Crowd}$, with the goal to assess whether ME leads to assessments of comparable quality to those from traditional scales. We compare ME and S_6 through the 95% confidence intervals of the mean truthfulness values for each grouping, using Tukey's test. Figure 3 shows the results. For both ME and S_6, all level pairs are significantly different in the GT_3 and GT_2 groupings.

When considering GT_6 directly, without grouping, the two scales show slight differences. For ME, Pants On Fire, False, and Mostly False (false levels) are not significantly different from each other, and the same applies to Half True, Mostly True, and True (true levels). However, all false levels differ significantly from all true levels. For S_6, Half True is statistically distinct from Mostly True and True, but not from Mostly False. In other words, ME differentiates false levels from Half True, while S_6 distinguishes Half True from the other true levels. While GT_6 effectively separates true from false information for ME, finer distinctions between adjacent levels are less apparent. Overall, the scales behave similarly, except for Half True.

We further compare ME and S_6 by computing the measures defined in Section 3.7. Table 3 shows the mean individual and aggregated scores. We assess statistical significance for each pair of scores, using a t-test for individual scores, a t-test for MAE and MSE, a binomial test for Accuracy, and a z-test for proportions for Pairwise. However, no tests are available to measure significance between Krippendorff's α scores for Internal and External.

When considering Accuracy, individual scores are slightly higher for S_6 than for ME, with statistically significant differences only

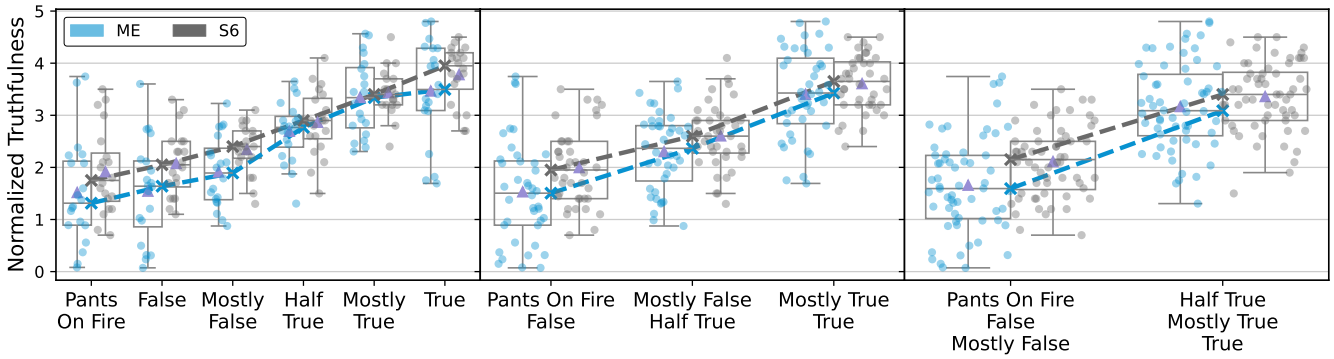


Figure 2: Comparison of aggregated and normalized ME_{Crowd} and $S6_{Crowd}$ assessments with ground truth (x-axis) for GT_6 (left), GT_3 (center), and GT_2 (right).

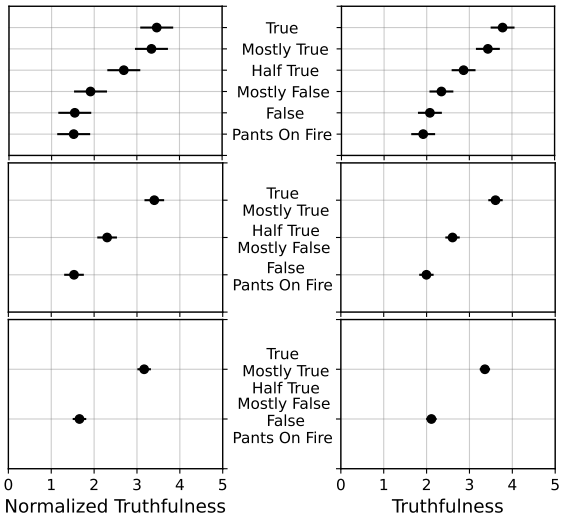


Figure 3: Comparison of 95% confidence intervals for mean aggregated truthfulness assessments computed using Tukey's test for each level of GT_6 (top), GT_3 (center), and GT_2 (bottom) groupings for ME (left) and S_6 (right).

for $Accuracy_3$ and $Accuracy_2$. Similar patterns are observed for MAE and MSE, where lower values indicate smaller error. For Pairwise, scores increase gradually across ground truth transformations, except for $Pairwise_2$, which is slightly higher for ME. Regarding agreement, both scales show similar levels of expert agreement (External). However, Internal is statistically higher for S_6 than for ME. Comparing Krippendorff's α scores between ordinal (S_6) and ratio (ME) scales may not be straightforward [45].

In summary, individual scores are higher for S_6 in $Accuracy_2$, $Accuracy_3$, $Accuracy_6$, MAE, MSE, External, and Internal, while for ME, they are higher in $Pairwise_2$, $Pairwise_3$, and $Pairwise_6$, with many of these differences being statistically significant. However, when aggregating scores, all measures increase compared to individual ones, and the differences are no longer statistically significant, suggesting comparable effectiveness between scales.

Table 3: Mean individual (left) and aggregated (right) effectiveness scores for $S6_{Crowd}$ and ME_{Crowd} assessments. † indicates a significant difference between S_6 and ME ($p < 0.05$); ‡ indicates $p < 0.01$; * indicates tests not possible.

	Individual		Aggregated	
Measure	**S_6**	**ME**	**S_6**	**ME**
$Accuracy_2$	0.65	0.63	0.83	0.80
$Accuracy_3$	0.51	$0.46^{\dagger}$	0.60	0.60
$Accuracy_6$	0.29	$0.24^{\ddagger}$	0.37	0.32
MAE	1.35	$1.66^{\ddagger}$	0.97	0.97
MSE	3.45	$4.83^{\ddagger}$	1.48	1.59
$Pairwise_2$	0.61	0.63	0.89	0.89
$Pairwise_3$	0.58	0.61	0.85	$0.83^{\ddagger}$
$Pairwise_6$	0.56	0.58	0.81	$0.79^{\ddagger}$
External	0.39	0.34	0.61	0.63^{*}
Internal	0.29	$0.15^{\ddagger}$	0.22	0.10^{*}

We now further investigate specific measures starting with External agreement. Figure 4 (first row) shows aggregated External agreement scores computed for each ground truth level and grouping. For the GT_6 grouping (top left plot), ME shows higher scores at each level, except for Half True. For the GT_3 grouping (top middle plot), ME shows higher scores for the first two levels, while S_6 shows higher scores for the true level. GT_2 grouping (top right plot) follows a similar trend.

We also perform bootstrapping (with 100 repetitions) using External scores. Figure 5 shows the distribution of aggregated External scores using $1 \leq n \leq 10$ assessments for each statement. Each dot represents a bootstrapped sample for which we compute the External score. Blue represents ME, while gray represents S_6. We use the t-test to compare differences between ME and S_6. For $1 \leq n \leq 3$ assessments, the External agreement computed for S_6 is statistically significantly higher than that for ME, indicating that S_6 provides better effectiveness when collecting this number of assessments. However, no statistically significant differences emerge between the scales for higher numbers of assessments. For both ME and S_6,

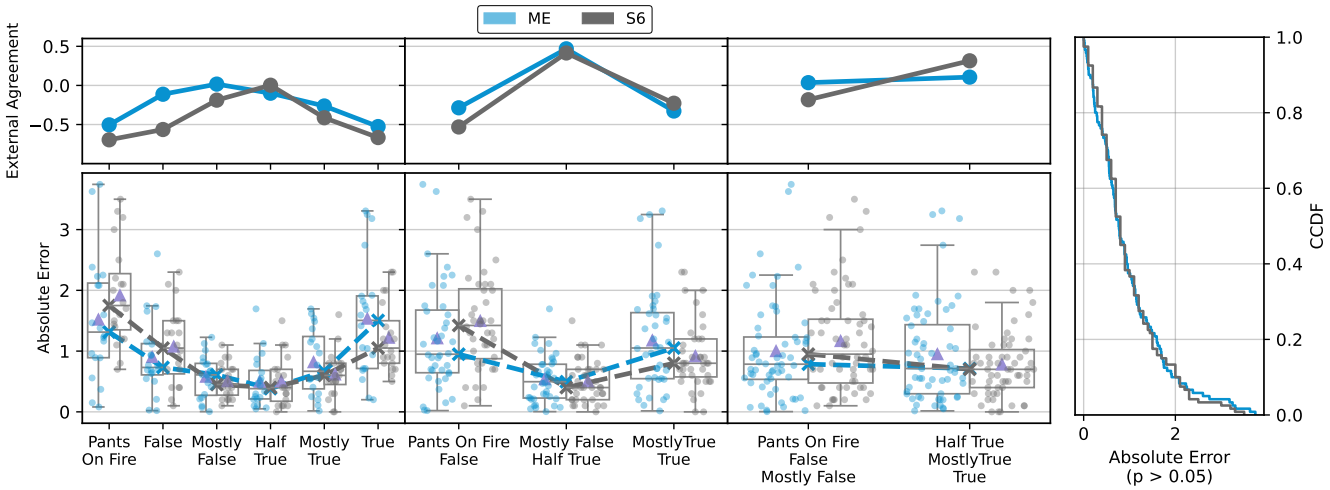


Figure 4: First row: Aggregated External agreement by ground truth level and scale for GT_6, GT_3, and GT_2 groupings. Second row: Absolute Error (AE) per statement based on aggregated truthfulness assessments for the same groupings. Right: Complementary Cumulative Distribution (CCDF) of AE values.

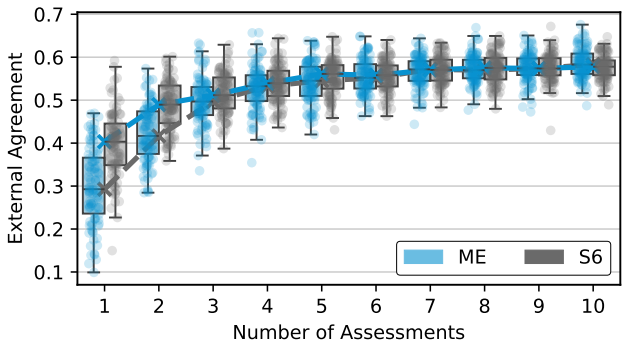


Figure 5: Bootstrapped External agreement scores with increasing assessments for ME (blue) and S_6 (gray).

increasing the number of assessments collected for each statement from $n = 7$ to $n = 10$ does not lead to a significant improvement in External. Thus, when using either ME or S_6 as an assessment scale, the number of assessments per statement can be reduced to 7 without compromising effectiveness, thereby decreasing the overall cost of the activity.

We now examine AE. Figure 4 (second row) shows the AE values for each statement's aggregated score, computed for each ground truth level and grouping. The x-axis represents the ground truth values, and the y-axis shows the aggregated AE values, with each circle representing a statement. For the GT_6 grouping (bottom left plot), we observe that S_6 exhibits higher error than ME for false ground truth levels, while ME shows higher error than S_6 for true levels. This pattern is consistent across GT_3 (bottom middle plot) and GT_2 (bottom right plot). Additionally, errors at the ends of the scales are larger than those at the center, which is expected due to the scale. Despite these differences, the AE values for each ground truth level across all groupings are not statistically significant, as shown by the CCDF of aggregated AE values (right plot). This indicates that the errors made by workers are similar for both scales, thus suggesting that the choice of scale does not influence the errors. A similar trend is observed with square errors.

We further investigate AE and SE for each scale by comparing the 95% confidence intervals of their mean values using Tukey's test (similar to Figure 3). For S_6, we observe statistically significant differences, as false ground truth levels exhibit higher errors than true ones. In contrast, for ME, no statistical significance is found.

In summary, when comparing ME with traditional scales such as S_6, there is no compelling reason to discard ME. The analysis of ME_{Crowd} and S_6 indicates that, overall, the two scales perform slightly differently, but the effectiveness of ME is comparable. Furthermore, the errors made by workers when assessing against the ground truth do not depend on the adopted scale, and both scales show similar degradation when fewer assessments are collected.

4.4 RQ3: ME Provides Additional Information

The results described in Sections 4.2 and 4.3 indicate that workers effectively use ME to provide truthfulness assessments. We now investigate how the truthfulness of information items is perceived at a finer level compared to traditional coarse-grained scales. We begin by investigating the gain profiles [45], which describe the perceived distances between truthfulness levels when comparing two assessment collections. Figure 6 shows the distributions of individual assessments for both the S_6 and ME scales, broken down by each ground truth value. For ME, we normalize the values. The black horizontal lines represent the aggregated scores: on the left, we aggregate using mean, and on the right, using median. The gray dashed lines connect the aggregated scores at the ends of the scales, representing a linear function, while the orange solid lines connect the aggregated scores for each level.

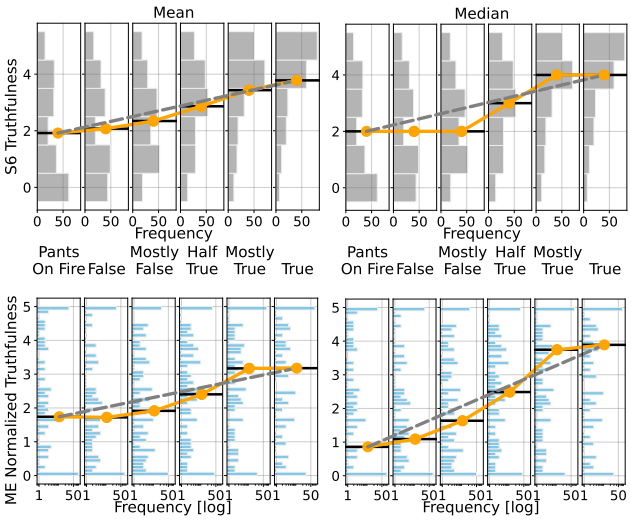


Figure 6: Perceived distances between assessments aggregated using mean (left) and median (right) for S_6 (top) and ME (bottom), with a breakdown for each ground truth level.

While it is reasonable to assume that workers perceive the six ground truth levels as equidistant, our assessments suggest otherwise. Workers do not perceive significant differences in truthfulness between information items evaluated at different levels by experts, such as Pants On Fire, False, and Mostly False, nor between Mostly True and True. This perception is stronger and more consistent with ME than with S_6, as the median aggregation function in S_6 limits detailed analysis. The median values can only reflect those of the scale, as illustrated in the top-right plot of the figure. Consequently, S_6 provides inconsistent signals depending on whether the median or mean is used for aggregation, making it inconclusive regarding the perception of distances between truthfulness levels. In contrast, ME consistently reflects the same perception, regardless of whether the median or mean is employed. When we consider the orange lines, each one, except for the S_6 mean (top left plot), resembles a sigmoid function centered at Half True, with lower and higher truthfulness levels perceived as more distinct.

In light of these results, we suggest an alternative approach to grouping GT_6 in contrast to previous studies, such as those by Roitero et al. [46] and La Barbera et al. [24]. Instead of combining GT_6 into GT_3, we propose a new GT_3 grouping that isolates Half True: Pants On Fire, False, and Mostly False; Half True; and Mostly True and True. We apply this updated GT_3 grouping and compare the 95% confidence intervals of mean truthfulness assessments for each level of the standard GT_3 with the updated version, using Tukey's test [60] for both ME and S_6 (similar to Figure 3). Although all levels of both versions of GT_3 are statistically significantly different from each other, the middle level, namely Mostly False + Half True and Half True, more distinctly separates adjacent levels in the updated version of GT_3. This revised version of GT_3 therefore leads to a trend that is closer to the commonly assumed linear one.

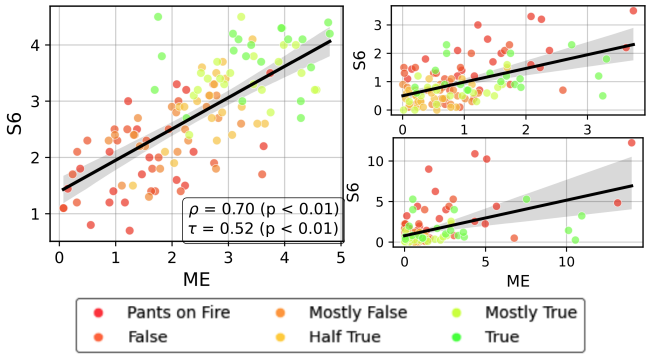


Figure 7: Correlation between the aggregated assessments (left), AE values (top right), and SE values (bottom right), provided using ME (x-axis) and S_6 (y-axis).

We turn to the correlation between the assessment collections. Figure 7 (left plot) shows the correlation between aggregated assessments of ME_{Crowd} and $S6_{Crowd}$, where each dot represents an individual statement, with red indicating false ground truth levels and green indicating true ones. A clear trend appears, with Pearson's ρ at 0.70 and Kendall's τ at 0.52. Similar correlations are observed for AE (top right plot) and SE (bottom right plot), with $\rho = 0.53$ and $\tau = 0.31$ for AE, and $\rho = 0.51$ and $\tau = 0.31$ for SE. These results suggest that although S_6 and ME perform similarly, the information items assessed correctly and not differ between the two scales. This is further supported by the Pairwise agreement between ME_{Crowd} and $S6_{Crowd}$, which is 0.75, indicating that the assessments agree on the ordering of two information items. These results imply that the two assessment collections provide complementary information and can be effectively combined.

5 Conclusions and Future Work

We presented the first systematic crowdsourcing study collecting truthfulness assessments using magnitude estimation (ME). Our findings show that ME aligns with expert fact-checkers (RQ1), performs comparably to traditional coarse-grained scales (RQ2), and captures more nuanced perceptions of truthfulness (RQ3).

In sum, ME emerges as a valuable alternative for truthfulness assessment, enhancing granularity and offering deeper insight into how information is perceived. As future work, we plan to analyze the assessments and the associated usability evaluations, investigate whether demographic and questionnaire data can improve aggregation, further explore ME's complementarity with other scales, and benchmark it against state-of-the-art automated fact-checking systems.

Acknowledgments

This research is partially supported by the PRIN 2022 Project – "MoT—The Measure of Truth: An Evaluation-Centered Machine-Human Hybrid Framework for Assessing Information Truthfulness" – Code No. 20227F2ZN3, CUP No. G53D23002800006 Funded by the European Union – Next Generation EU – PNRR M4 C2 I1.1.

References

[1] Jennifer Allen, A. Arechar, Gordon Pennycook, and G. David Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7, 36 (2021), 10 pages. https://doi.org/10.1126/sciadv.abf4393

[2] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 245, 19 pages. https://doi.org/10.1145/3491102.3502040

[3] Enrique Amigó, Julio Gonzalo, and Stefano Mizzaro. 2023. What is My Problem? Identifying Formal Tasks and Metrics in Data Mining on the Basis of Measurement Theory. *IEEE Transactions on Knowledge and Data Engineering* 35, 2 (2023), 2147–2157. https://doi.org/10.1109/TKDE.2021.3109823

[4] Enrique Amigó, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de Albornoz. 2020. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3938–3949. https://doi.org/10.18653/v1/2020.acl-main.363

[5] Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude Estimation of Linguistic Acceptability. *Language* 72, 1 (1996), 32–68. https://doi.org/10.2307/416793

[6] Erik Brand, Kevin Roitero, Michael Soprano, Afshin Rahimi, and Gianluca Demartini. 2022. A Neural Model to Jointly Predict and Explain Truthfulness of Statements. *Journal of Data and Information Quality* 15, 1, Article 4 (12 2022), 19 pages. https://doi.org/10.1145/3546917

[7] Sijing Chen, Lu Xiao, and Akit Kumar. 2023. Spread of misinformation on social media: What contributes to it and how to combat it. *Computers in Human Behavior* 141 (2023), 107643. https://doi.org/10.1016/j.chb.2022.107643

[8] Ralph B. D'Agostino, Anastasios Belanger, and Ralph B. D'Agostino Jr. 1990. A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician* 44, 4 (1990), 316–321. https://doi.org/10.1080/00031305.1990.10475751

[9] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE* 18, 3 (03 2023), 1–17. https://doi.org/10.1371/journal.pone.0279720

[10] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2114–2124. https://doi.org/10.1145/3531146.3534629

[11] Trygg Engen and Donald H McBurney. 1964. Magnitude and category scales of the pleasantness of odors. *Journal of Experimental Psychology* 68, 5 (1964), 435–440. https://doi.org/10.1037/h0041008

[12] Sam Featherston. 2005. Universals and grammaticality: wh-constraints in German and English. *Linguistics* 43, 4 (2005), 667–711. https://doi.org/10.1515/ling.2005.43.4.667

[13] D Fucci, L Petrosino, D McColl, D Wyatt, and C Wilcox. 1997. Magnitude estimation scaling of the loudness of a wide range of auditory stimuli. *Perceptual and Motor Skills* 85, 3 Pt 1 (12 1997), 1059–1066. https://doi.org/10.2466/pms.1997.85.3.1059

[14] Shin Fukuda, Grant Goodall, Dan Michel, and Henry Beecher. 2012. Is Magnitude Estimation Worth the Trouble?. In *Proceedings of the 29th West Coast Conference on Formal Linguistics*. Cascadilla Proceedings Project, Somerville, MA, 328–336. https://www.lingref.com/cpp/wccfl/29/paper2718.pdf Document #2718.

[15] George A. Gescheider. 1997. *Psychophysics: The Fundamentals* (3rd ed.). Psychology Press, Hove, East Sussex, United Kingdom. https://doi.org/10.4324/9780203774458

[16] Lucas Graves and Michelle Amazeen. 2019. Fact-Checking as Idea and Practice in Journalism. In *Oxford Research Encyclopedia of Communication*, J. Nussbaum (Ed.). Oxford University Press, Oxford, United Kingdom. https://doi.org/10.1093/acrefore/9780190228613.013.808

[17] Tom L. Han, Kevin Roitero, Ujwal Gadiraju, C. Sarasua, A. Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The Impact of Task Abandonment in Crowdsourcing. *IEEE Transactions on Knowledge & Data Engineering* 1, 1 (10 2019), 1–1. https://doi.org/10.1109/TKDE.2019.2948168

[18] David J. Hauser, Aaron J. Moss, Cheskie Rosenzweig, Shalom N. Jaffe, Jonathan Robinson, and Leib Litman. 2023. Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods* 55, 8 (01 Dec 2023), 3953–3964. https://doi.org/10.3758/s13428-022-01999-x

[19] Simon Kemp. 1991. Magnitude Estimation of the Utility of Public Goods. *Journal of Applied Psychology* 76, 4 (1991), 533–540. https://doi.org/10.1037/0021-9010.76.4.533

[20] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. 2020. The Shape of And Solutions To The MTurk Quality Crisis. *Political Science Research and Methods* 8, 4 (2020), 614–629. https://doi.org/10.1017/psrm.2020.6

[21] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. 2020. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods* 8, 4 (2020), 614–629. https://doi.org/10.1017/psrm.2020.6

[22] Frederick A.A. Kingdom and Nicolaas Prins. 2016. Chapter 3 - Varieties of Psychophysical Procedures. In *Psychophysics (Second Edition)* (second edition ed.), Frederick A.A. Kingdom and Nicolaas Prins (Eds.). Academic Press, San Diego, 37–54. https://doi.org/10.1016/B978-0-12-407156-8.00003-7

[23] Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. *UPENN Libraries* 1 (2011), 43. https://repository.upenn.edu/asc_papers/43

[24] David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina, and Stefano Mizzaro. 2024. Crowdsourced Fact-checking: Does It Actually Work? *Information Processing & Management* 61, 5 (2024), 103792. https://doi.org/10.1016/j.ipm.2024.103792

[25] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II* (Lisbon, Portugal). Springer-Verlag, Berlin, Heidelberg, 207–214. https://doi.org/10.1007/978-3-030-45442-5_26

[26] David La Barbera, Michael Soprano, Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro. 2023. Fact-Checking at Scale with Crowdsourcing: Experiments and Lessons Learned. In *Proceedings of the 13th Italian Information Retrieval Workshop (CEUR Workshop Proceedings, Vol. 3448)*. CEUR-WS.org, Pisa, Italy, 85–90. https://ceur-ws.org/Vol-3448/paper-18.pdf

[27] Guoliang Li, Jiannan Wang, Yudian Zheng, Ju Fan, and Michael J. Franklin. 2018. *Crowdsourced Data Management: Hybrid Machine-Human Computing*. Springer Singapore, Singapore. https://doi.org/10.1007/978-981-10-7847-7

[28] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2015. Judging Relevance Using Magnitude Estimation. In *Advances in Information Retrieval*. Springer International Publishing, Cham, 215–220. https://doi.org/10.1007/978-3-319-16354-3_23

[29] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Transactions on Information Systems* 35, 3, Article 19 (2017), 32 pages. https://doi.org/10.1145/3002172

[30] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/3121050.3121060

[31] Lawrence E. Marks. 1988. Magnitude estimation and sensory matching. *Perception & Psychophysics* 43, 6 (01 Dec 1988), 511–525. https://doi.org/10.3758/BF03207739

[32] Catherine C. Marshall, Partha S.R. Goguladinne, Mudit Maheshwari, Apoorva Sathe, and Frank M. Shipman. 2023. Who Broke Amazon Mechanical Turk? An Analysis of Crowdsourcing Data Quality over Time. In *Proceedings of the 15th ACM Web Science Conference 2023* (Austin, TX, USA) *(WebSci '23)*. Association for Computing Machinery, New York, NY, USA, 335–345. https://doi.org/10.1145/3578503.3583622

[33] Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G. Rand. 2024. Crowds Can Effectively Identify Misinformation at Scale. *Perspectives on Psychological Science* 19, 2 (2024), 477–488. https://doi.org/10.1177/17456916231190388

[34] S. McDaniel, J. D. Leatherwood, and B. M. Sullivan. 1992. *Application of magnitude estimation scaling to the assessment of subjective loudness response to simulated sonic booms*. Technical Report NASA-TM-107657. NASA. https://core.ac.uk/download/pdf/42811384.pdf

[35] Mick McGee. 2003. Usability Magnitude Estimation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 47, 4 (2003), 691–695. https://doi.org/10.1177/154193120304700406

[36] Paula De Pando Mena. 2019. Principles and Boundaries of Fact-checking: Journalists' Perceptions. *Journalism Practice* 13, 6 (2019), 657–672. https://doi.org/10.1080/17512786.2018.1547655

[37] Howard R. Moskowitz. 1977. Magnitude Estimation: Notes on What, How, When, and Why to Use It. *Journal of Food Quality* 1, 3 (1977), 195–227. https://doi.org/10.1111/j.1745-4557.1977.tb00942.x

[38] Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Kristina Toutanova and Hua Wu (Eds.). Association for Computational Linguistics, Baltimore, Maryland, 1009–1019. https://doi.org/10.3115/v1/P14-1095

[39] Feng Ni, David Arnott, and Shijia Gao. 2019. The Anchoring Effect In Business Intelligence Supported Decision-making. *Journal of Decision Systems* 28, 2 (2019), 67–81. https://doi.org/10.1080/12460125.2019.1620573

[40] Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526. https://doi.org/10.1073/pnas.1806781116

[41] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 2173–2178. https://doi.org/10.1145/2983323.2983661

[42] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 231–240. https://doi.org/10.18653/v1/P18-1022

[43] Aaron Rich and Mick McGee. 2004. Expected Usability Magnitude Estimation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 48, 5 (2004), 912–916. https://doi.org/10.1177/154193120404800536

[44] Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2018. How Many Truth Levels? Six? One Hundred? Even More? Validating Truthfulness of Statements via Crowdsourcing. In *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management*. CEUR-ws.org, Torino, Italy, 1–6. http://ceur-ws.org/Vol-2482/paper38.pdf

[45] Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. *Information Processing & Management* 58, 6 (2021), 102688. https://doi.org/10.1016/j.ipm.2021.102688

[46] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43st International ACM SIGIR Conference on Research and Development in Information Retrieval*. (Xi'an, China (Virtual)) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 439–448. https://doi.org/10.1145/3397271.3401112

[47] Kevin Roitero, Michael Soprano, David La Barbera, Eddy Maddalena, and Stefano Mizzaro. 2024. Enhancing Fact-Checking: From Crowdsourced Validation to Integration with Large Language Models. In *Proceedings of the 14th Italian Information Retrieval Workshop (CEUR Workshop Proceedings, Vol. 3802)*. CEUR-WS.org, Udine, Italy, 74–77. https://ceur-ws.org/Vol-3802/paper13.pdf

[48] Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2023. Can The Crowd Judge Truthfulness? A Longitudinal Study On Recent Misinformation About COVID-19. *Personal and Ubiquitous Computing* 27, 1 (1 2 2023), 59–89. https://doi.org/10.1007/s00779-021-01604-6

[49] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020. The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1305–1314. https://doi.org/10.1145/3340531.3412048

[50] Jacques Rutschmann. 1973. Time Judgments by Magnitude Estimation and Magnitude Production and Anxiety: A Problem of Comparison Between Normals and Certain Schizophrenic Patients. *The Journal of Psychology* 85, 2 (1973), 187–223. https://doi.org/10.1080/00223980.1973.9915649

[51] Falk Scholer, Eddy Maddalena, Stefano Mizzaro, and Andrew Turpin. 2014. Magnitudes of Relevance: Relevance Judgements, Magnitude Estimation, and Crowdsourcing. In *Proceedings of the Sixth International Workshop on Evaluating Information Access*. National Institute of Informatics (NII), Tokyo, Japan, 9–16. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/EVIA/02-EVIA2014-ScholerF.pdf

[52] Vivian Shinobu, Utamura V Chikako, Murase V Yukiyo, Hamatani, and Yukinori Nagano. 2009. User Experience Index Scale: Quantifying Usability by Magnitude Estimation. *Fujitsu Scientific & Technical Journal* 45 (2009), 219–225. https://www.fujitsu.com/global/documents/about/resources/publications/fstj/archives/vol45-2/paper05.pdf

[53] Michael Soprano, Kevin Roitero, Francesco Bombassei De Bona, and Stefano Mizzaro. 2022. Crowd_Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks Off-the-shelf. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1605–1608. https://doi.org/10.1145/3488560.3502182

[54] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. 2024. Cognitive Biases in Fact-Checking and Their Countermeasures: A Review. *Information Processing & Management* 61, 3 (2024), 103672. https://doi.org/10.1016/j.ipm.2024.103672

[55] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale. *Information Processing & Management* 58, 6 (2021), 102710. https://doi.org/10.1016/j.ipm.2021.102710

[56] Antonella Sorace and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115, 11 (2005), 1497–1524. https://doi.org/10.1016/j.lingua.2004.07.002 Data in Theoretical Linguistics.

[57] S. S. Stevens. 1951. Mathematics, Measurement, and Psychophysics. In *Stevens' Handbook of Experimental Psychology*. Wiley, Oxford, England, 1–49. https://doi.org/10.1002/9781119170174

[58] S. S. Stevens. 1975. *Psychophysics: Introduction to its perceptual, neural, and social prospects*. John Wiley & Sons, Oxford, England. v, 329–v, 329 pages. https://awspntest.apa.org/record/1975-20087-000

[59] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[60] John W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA. https://archive.org/details/exploratorydataa00tuke

[61] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 2015. The Benefits of Magnitude Estimation Relevance Assessments for Information Retrieval Evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 565–574. https://doi.org/10.1145/2766462.2767760

[62] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17, 3 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[63] Virtanen, Pauli and others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html. Accessed: 2025-04-28.

[64] Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics, Baltimore, MD, USA, 18–22. https://doi.org/10.3115/v1/W14-2508

[65] Ziying Yang, Alistair Moffat, and Andrew Turpin. 2018. Pairwise Crowd Judgments: Preference, Absolute, and Ratio. In *Proceedings of the 23rd Australasian Document Computing Symposium* (Dunedin, New Zealand) *(ADCS '18)*. Association for Computing Machinery, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/3291992.3291995