Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty

QUANZE CHEN, University of Washington, USA
DANIEL S. WELD, University of Washington, USA
AMY X. ZHANG, University of Washington, USA

Human ratings have become a crucial resource for training and evaluating machine learning systems. However, traditional elicitation methods for absolute and comparative rating suffer from issues with consistency and often do not distinguish between uncertainty due to disagreement between annotators and ambiguity inherent to the item being rated. In this work, we present Goldilocks, a novel crowd rating elicitation technique for collecting calibrated scalar annotations that also distinguishes inherent ambiguity from inter-annotator disagreement. We introduce two main ideas: grounding absolute rating scales with examples and using a two-step bounding process to establish a range for an item's placement. We test our designs in three domains: judging toxicity of online comments, estimating satiety of food depicted in images, and estimating age based on portraits. We show that (1) Goldilocks can improve consistency in domains where interpretation of the scale is not universal, and that (2) representing items with ranges lets us simultaneously capture different sources of uncertainty leading to better estimates of pairwise relationship distributions.

CCS Concepts: • **Human-centered computing → Collaborative interaction**.

Additional Key Words and Phrases: crowdsourcing; annotation; ambiguity; calibration

ACM Reference Format:
Quanze Chen, Daniel S. Weld, and Amy X. Zhang. 2021. Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 335 (October 2021), 25 pages. https://doi.org/10.1145/3476076

1 INTRODUCTION

Much of modern machine learning is built on the foundation of human-annotated data. As the application of these models has expanded into more socially embedded and contextually nuanced domains [2, 51, 52], collecting high quality, consistent, and robust data from human annotators has become an increasingly important yet challenging task [6]. As one example, the ability to gather human evaluations of the toxicity of a piece of text is a necessary precursor to being able to build toxicity models to support online communities [70] as well as capture and mitigate harmful outputs generated by large language models [28].

However, traditional rating methods commonly used today, like absolute or comparative rating, can produce inconsistencies in ratings across annotators and even with a single annotator's ratings [3, 57]. This is due to issues such as lack of a common interpretation of the scale in the case of absolute rating, as well as lack of global context in the case of comparative rating [17, 68, 69].

Authors' addresses: Quanze Chen, cqz@cs.washington.edu, University of Washington, Seattle, WA, USA; Daniel S. Weld, weld@cs.washington.edu, University of Washington, Seattle, WA, USA; Amy X. Zhang, axz@cs.washington.edu, University of Washington, Seattle, WA, USA.

2573-0142/2021/10-ART335 $15.00
https://doi.org/10.1145/3476076

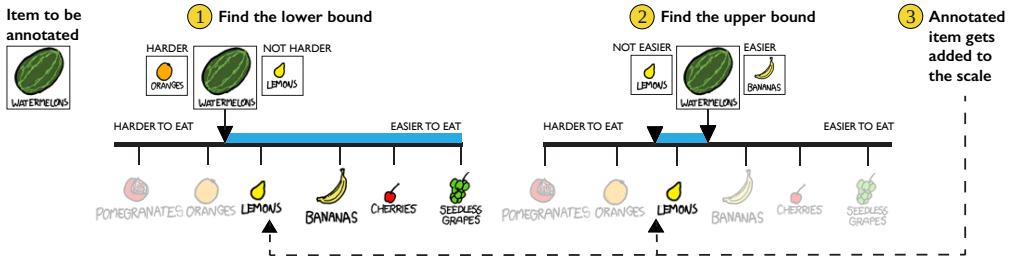


Fig. 1. The Goldilocks annotation process involves placing items onto a continuous scale that is populated with items that have previously been annotated. The process is broken down into three parts. (1) Find the lower bound by moving the left handle of the slider towards the right and away from its initial position on the far left of the scale. Continue until encountering an item on the scale that is either greater than or indistinguishable from the item to be annotated. (2) Find the upper bound in the same way but moving the right handle towards the left. Continue until encountering an item on the scale that is less than or indistinguishable from the item to be annotated or until the two handles are on top of each other, representing complete certainty. (3) Finally, the lower and upper bounds of the item get added to the scale to join the existing items. Thus, an annotator will be able to see and compare against their own prior annotated items as they annotate more items. Images of fruit are taken from XKCD: https://xkcd.com/388/

Additionally, while current rating methods can capture uncertainty in the ratings, it is difficult to dissect whether the uncertainty is a result of inherent ambiguity in the item—where certain items cannot be confidently distinguished from each other [23]—or from disagreement between annotators on where the item should be placed. Distinguishing these sources of uncertainty offers the potential of better capturing biases between annotators. It also allows us to develop more calibrated models that only make high-confidence distinctions between items when a human would have as well [33].

In this paper, we propose a new design for collecting scalar annotations called Goldilocks[1] that combines the ability to make direct comparisons between items with the simplicity of a continuous absolute rating scale (Figure 1). To accomplish this, Goldilocks introduces two main ideas—(1) *Calibration using Prior Annotations*: we provide previously annotated items as anchors to ground interpretations of the scale both within and across annotators. (2) *Item-level Resolution Elicitation using Ranges*: we use a two-step process to collect lower and upper bounds for each item instead of a single placement. Goldilocks combines strengths from both absolute and comparative ratings as annotators make multiple comparative judgments while placing an item on an absolute scale. In addition, by directly eliciting an annotator's own judgment of an item's inherent ambiguity instead of relying on aggregating inter-annotator agreement, Goldilocks can separate agreement from perceived ambiguity.

To understand the effectiveness of these designs, we conducted three studies comparing aspects of the Goldilocks annotation process against traditional methods. In the first experiment, we evaluated whether anchoring scales with a shared set of previously annotated items can improve consistency of item placement across annotators. In the second experiment, we examined whether including an annotator's own prior annotations as anchors improves self-consistency. Our final experiment evaluated how well ranges captured using Goldilocks can recover the distribution of pairwise relationships as measured by traditional absolute and comparative rating. Each of our experiments were conducted in three domains representative of the subjective or ambiguous rating tasks that can be challenging for traditional methods: judging TOXICITY of online comments (short

[1]Somewhat like Goldilocks in "Goldilocks and the Three Bears", annotators must make use of *multiple* comparisons.

text), estimating SATIETY of food depicted in images (visual), and estimating AGE from portrait photos (visual).

From the experiments examining anchors, we found that the addition of shared example anchors to ground rating scales improves rating consistency between annotators in domains where shared understanding of the scale is low. We also found indications that showing one's prior annotations in a session as additional anchors may improve self-consistency on examples where there is high initial uncertainty. From the experiment examining ranges, we found that our two-step range annotation process allows us to infer pairwise relationship distributions that are more robust—simultaneously reflecting both uncertainty of single annotators and disagreement between annotators—compared to alternatives with a single value. Finally, we found that the size of range annotations provides an interpretation of uncertainty that is distinct from the uncertainty modeled via inter-annotator disagreement.

We conclude with a discussion of the limitations and opportunities for Goldilocks. Regarding efficiency, while our approach is more costly than performing just one of absolute or comparative rating, our method is cheaper than performing both, which would be necessary to recover the richer data that Goldilocks generates. We discuss cases where a deeper understanding of uncertainty can be important for generating more trustworthy model predictions. We also discuss what we envision as a scaled-up Goldilocks workflow: utilizing iterative improvement through multiple annotation sessions with designs for bootstrapping the initial set of anchors along with interesting problems to be explored in each of these aspects.

2 RELATED WORK

In this section we review prior work on: (1) growing demand for consistent and robust human rating, (2) prior work building on absolute and comparative rating designs, (3) uncertainty and disagreement in crowd annotation, and (4) making use of uncertainty from human annotators in downstream machine learning tasks.

2.1 Demand for Improving Human Rating

There is a growing demand for human annotation in domains involving ambiguous or subjective examples, largely due to rapid progress in machine learning. Human rating annotation has been used to create or validate a variety of training data, for example, in the domains investigating toxicity [70], misinformation and credibility [6, 51], and emotionally manipulative text [35]. However, there is also increasing concern for the robustness of datasets collected [69] and whether nuances like uncertainty are being represented [4].

Direct human rating of model output has also become prevelant in the evaluation of high performance models where automated metrics (e.g., BLEU, METEOR) start to fail [1, 12, 19]. For example, human rating has been used to evaluate aspects of generative tasks (e.g., summarization, translation) in natural language processing by capturing characteristics like fluency, relevance, and conciseness which cannot be easily and reliably assessed with automated metrics [32]. Human rating has also commonly been used to evaluate the output of chatbots [59] or to judge search results [36] or cluster quality [74]. Increasingly, human ratings (both comparative and absolute) are becoming an integral aspect in facilitating comparisons between models through evaluation leaderboards and shared tasks [41, 61], where consistency and robustness of comparative results are crucial.

2.2 Absolute and Comparative Rating Designs

One of the most common designs for collecting human ratings today is through *absolute rating* scales, often in the form of Likert or semantic differential scales [46, 53]. When a consistent

interpretation of the scale can be established across annotators, designs based on absolute rating can offer many benefits such as being very efficient (only requiring a single annotation per item) and providing easily interpretable ratings that are globally contextualized (rather than depending on other items). However, many annotation domains do not have commonly accepted scales, meaning that divergent interpretations of a scale based on abstract text descriptions can become a source of disagreement and inconsistency across annotators [68]. Even within an annotator's own annotations, the lack of a well defined scale means that to maintain consistent ratings, they must refer to their own memory of their past decisions which can be unreliable [10]. Accounting for these inconsistencies requires additional effort—either through additional calibration [27] or just identifying and reporting them [29]. Absolute scales can also be locally unreliable [69]—because items are only ever compared against the scale's anchors, pairwise comparisons between two items with similar values can only be rigorously done if the measurement resolution (uncertainty around the values) is also accounted for.

As many consistency problems in absolute rating systems result from the lack of direct comparisons between actual items, a natural solution is to look towards the other major alternative—*comparative ratings* [64]. In comparative rating systems, items are compared against against one another directly, circumventing the need for a scale as a proxy and providing highly reliable measurements of local relationships. This kind of comparison can also be more intuitive for annotators leading to comparative systems sometimes suggested as a more accurate alternative for ranking items [42, 45]. However, collecting comparative ratings can be considerably more costly (on the order of N comparisons per item) unless sampling and ranking aggregation methods or partial comparisons, which trade off additional uncertainty, are used [38, 42]. The focus on local comparisons makes it easy for an annotator to inadvertently produce annotations that are not globally self-consistent, requiring post-hoc corrective action that may not reflect an annotator's actual judgment. Abandoning global context also means that if a rating score (rather than ranking) is desired, a numeric mapping like Elo rating needs to be done [17], which often come with assumptions about uniform spacing between items.

Past work has explored hybrid approaches that combine aspects of comparative and absolute annotation. For example, Sakaguchi et al. [56] present EASL, a hybrid approach where items are rated using continuous absolute scales but similar items are grouped together for annotation allowing for some degree of comparison and contextualization. While similar in motivation, our work differs in that we make comparison an integral part of the annotation process rather than an optional source of context, allowing us to provide more consistency by grounding comparison against global anchors and capture uncertainty intuitively by using comparisons to establish bounds.

Beyond the individual drawbacks mentioned above, neither of the two traditional annotation methods supports effective separation of the sources of uncertainty as a part of the the annotation process [37]. These sources include both aleatoric uncertainty, or irreducible ambiguity inherent to the item being rated, and epistemic uncertainty, or disagreement on the placement of the item. Absolute rating forces annotators to resolve inherent ambiguity into a precise placement causing both sources of uncertainty to be mixed. Meanwhile, comparative rating only provides an indirect view into inherent ambiguity through the size of equivalence sets. Separating the two sources of uncertainty is especially desirable as it can be an important tool for understanding properties of the items being annotated separate from biases or divergent interpretations among annotators.

2.3 Addressing Uncertainty and Disagreement

Uncertainty and disagreement has been a long recognized challenge when collecting crowdsourced annotations of all kinds. Early work in crowdsourcing focused on measuring perception-based objective aspects of items, taking the view that the uncertainty observed as disagreement between

workers is the result of random noise from unreliable perception. To address this kind of uncertainty, methods such as majority voting, expectation maximization [18], Max-Margin Majority Voting [65] and even active learning based approaches [48] have been proposed, which attempt to improve the quality of the true measurement signal by aggregating across more annotators and accounting for the varying degree of noise introduced by each annotator. More recent lines of work recognize the deficiencies in single value answers, proposing instead to use answer distributions in the form of allowing multiple labels [22, 23, 39] to capture the sources of uncertainty more comprehensively rather than attempt to remove it. Generally, these aggregation methods rely on large amounts of redundancy; thus, a major focus of prior work in this area has been in improving the efficiency of annotation work though collecting more information for each item or information about more items in each annotation task [16, 42].

Another view focuses on the idea that disagreements can arise from divergent interpretations of the data and task specification among workers [4, 31, 40] or even within an annotator as they are exposed to more data. One prior line of work, structured labeling [13, 43], proposes that tools and techniques can be designed to assist people in reconciling the evolving interpretations of data both individually and collectively when labeling or generating taxonomies. Kuleza et al. [43] note that maintaining consistency in annotation can be challenging even for experts. This motivated our exploration of improving self-consistency by incorporating past annotations.

Rubrics [73] and training (such as via gated instructions [50]) have also been proposed as an effective way to unify understanding of the task requirements across workers. In fact, in practice, task designers oftentimes expend significant effort building detailed rubrics with complex training and gating processes. Prior work in this area has explored how to reduce such burdens on task designers through collaboratively creating rubrics with workers [9]. However, strict rubrics are often undesirable when the goal is to elicit human judgments on properties that are difficult to rigorously define such as those involving subjective interpretation.

A softer form of rubrics can be made by using in-domain ground truth (or "gold") examples to anchor the interpretation of tasks including those involving scales. Gold solutions created by experts are often used during training [20] in lieu of or in addition to instructions and rubrics. Reference examples can also be provided during the task such as in the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA test) [66]. However, existing methods depend on curated or synthesized fixed gold anchors ahead of labeling, which, in the case of scale anchors for subjective domains, still requires a concrete definition of the scale ahead of time. Fixed anchors are also limited in the support they can provide for self-consistency.

Finally, on challenging or high-stakes domains where correctness is important, deliberation has been proposed as a way of addressing and resolving disagreement directly. Deliberation processes can range from simple one-shot reconsideration prompts [21] to more complex multi-turn discussions [15, 58]. Alternatively, lighter methods have been proposed that model behavior of humans to identify when disagreement is likely [34]. However, these methods can still require significant worker effort.

2.4 Recognizing and Leveraging Uncertainty in ML

As it is impossible to eliminate uncertainty [37], downstream tasks like machine learning have started to explore paths of utilizing uncertainty information (when it is available) during training and evaluation. Many machine learning models have been built to do tasks like classification [8], ranking [71] or regression [72] using uncertain labels.

Recently, there is growing interest in understanding and mitigating adversarial attacks on machine learning [30]. These attacks often trick models to make high confidence predictions that are incorrect due to limited ability for many models to accurately model its own confidence. One

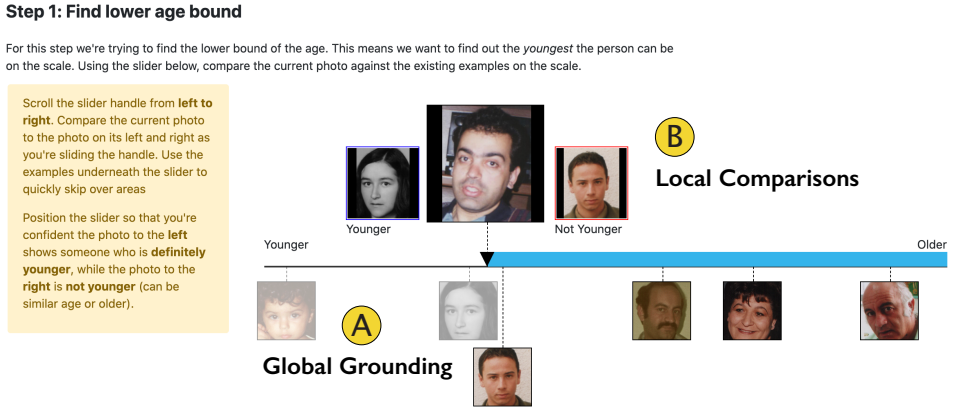


Fig. 2. A screenshot showing the comparisons that annotators can make while placing the upper or lower bound of an item on the scale in the Goldilocks annotation process. To support grounding with examples, Goldilocks provides: (A) global grounding by selecting 5–7 previously annotated items that are maximally spread out on the scale and placing them as anchors to support coarse and fast global adjustment. (B) Local comparisons of previously annotated items directly to the left and right of the slider handle are shown as an annotator scrubs the handle across the slider. Local items that are not one of the global examples are inserted as anchors. Together, this allows annotators to make fine-grained local adjustments.

potential mitigation strategy has been to look at improving a model's robustness by improving ability to model uncertainty [54]. Additionally, there is a push for models to more accurately understand when it should be unsure [55]. These all motivate an increasing need to understand what humans find uncertain and calibrated ways for humans to convey their degree of uncertainty.

3 DESIGN

Absolute rating can suffer from inconsistent scale interpretations while comparative rating lacks global context. Our design for the Goldilocks annotation system takes a hybrid approach, with the specific goals of: (1) improving consistency (between annotators and over time within annotator), and (2) enabling intuitive indication of uncertainty with respect to the scale for each example being labeled.

In this section, we will describe the designs that address each of the goals above followed by additional aspects of operating the complete annotation workflow. At the end, we will discuss specific details of the design decisions we made for our implementation separate from the overall design of the Goldilocks annotation process.

3.1 Grounding with Prior Examples

We base the main interactions in Goldilocks around an absolute rating design. To mitigate the aforementioned drawbacks of absolute rating, Goldilocks uses prior examples in addition to abstract descriptions to ground the scale, making it possible to make pairwise comparisons while still using absolute rating interactions. Prior work has shown that human judgments measured explicitly with comparisons can be easier than direct labels for some tasks [60, 67, 75], and *fixed* reference anchors have been used in other procedures to provide a more concrete grounding of scales [66]. Similar ideas that use comparisons against samples to contextualize abstract scales also exist in other fields like cognitive psychology [63].

Goldilocks uses a set of previously-annotated examples to add two additional pieces of information to the absolute rating scale—**global grounding** and **local comparisons**, as shown in Figure 2. With **global grounding**, a small set of representative examples are selected and placed as anchors along the rating scale, similar to existing text-based anchors for levels in traditional absolute rating. Using concrete examples allows annotators to quickly understand and estimate where each item could fit on the scale. Since there can be many previously-annotated examples, we make sure to only visualize a smaller subset of examples (around 5 to 7, similar to typical numbers of Likert levels) that are maximally spread out along the scale. In practice, there are many ways to select these examples. The specific selection process we used is outlined in 3.4.

While global grounding is useful for making coarse placements, it alone is insufficient for narrowing down specific placement of items. To help the annotators find specific placements, Goldilocks also surfaces **local comparisons** by showing the immediate neighborhood above and below a position on the scale. As annotators scrub along a continuous scale, we show side-by-side comparisons between the current indicated position and the closest items above and below this position. Placements of these neighbors are also indicated on the scale itself, allowing for annotators to adjust proportional distance to each neighbor based on their evaluation of the item being placed. These designs together allow for a more consistent and concrete instantiation of the scale across multiple annotators.

Finally, Goldilocks addresses local self-consistency by supporting dynamic augmentation of the anchor examples used to ground the rating scale: as annotators progress in an annotation session, their own annotations for earlier items are also incorporated into the set of references alongside any pre-seeded ones (Step 3 of Figure 1). These personal annotations will then also take part in both global grounding and local comparisons, making it possible to directly compare new items against past annotations produced in the same session.

One potential limitation for any annotation process involving examples is how to start the annotation when no past examples are available. Goldilocks accounts for this with a separate procedure to curate an initial seed set that is deployed when past examples do not exist. We will dive into more detail about the selection of this initial seed set of items to jumpstart annotation in Section 3.3. In the discussion section, we will also discuss avenues of addressing other challenges in example-based grounding such as scaling up annotation with iterative improvement and addressing density as the scale becomes populated with more annotated examples.

3.2 Two-Step Range Annotation

Not all items can be meaningfully distinguished from all other items by an annotator. Instead of forcing the breaking of ties, most designs for side-by-side comparisons allow annotators to indicate "indistinguishable" or "tied" pairs [44]—however, there is no such elicitation process for traditional absolute rating designs. With Goldilocks, we propose a new process that allows annotators to indicate "indistinguishable" pairwise relationships on an absolute rating scale. To achieve this, we take inspiration from prior work [24], where annotators were asked to select *all* potentially relevant labels for an item instead of a single best label option. We extend this into the continuous scale domain by introducing the concept of eliciting "range" labels—where upper and lower bounds establish a subsection of the scale representing where an item *could* be placed. Our range-based approach is also reminiscent of methods like best-worst scaling [42] in comparative rating, which can efficiently capture pairwise relationships across many items.

Prior designs have explored alternatives to eliciting uncertainty for scalar annotations, such as in the form of a weighted distribution across *surrounding* anchor labels [16]. However, estimating distributions in this way can be challenging for humans, as an annotator has little guidance on how to allocate weight to the anchoring labels they find reasonable. In Goldilocks, we can

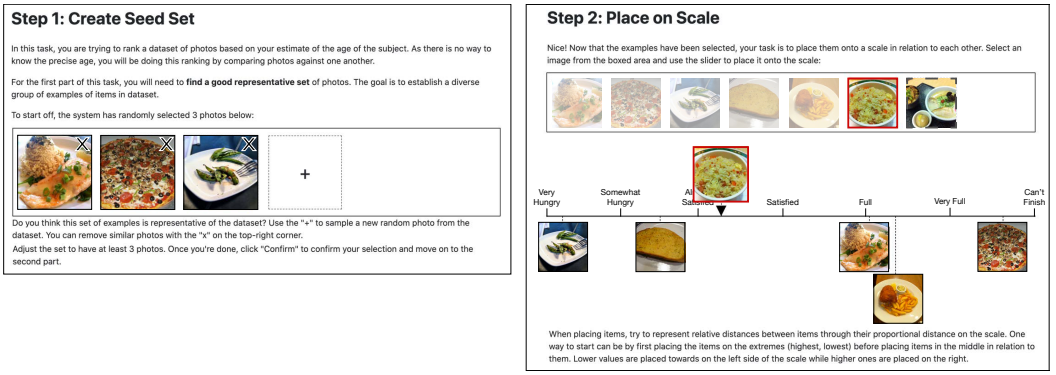


Fig. 3. Screenshots illustrating the two steps in the cold start process for Goldilocks. Step 1 (Left): A seed set can be created by using the cold start interface to randomly draw examples and drop existing ones to create an adequately sized representative set of examples. Step 2 (Right): The items from the seed set are placed onto a scale by adjusting their position relative to each other, forming the initial values that can be used to bootstrap the annotation tasks in Goldilocks. These initial items can later be reintroduced in the Goldilocks annotation process once other items have been annotated, in order to collect ranges.

take advantage of the comparisons afforded by grounding examples to contextualize distributions intuitively. Specifically, we break down the process of eliciting ranges into two steps: finding the lower bound and then finding the upper bound (Steps 1 and 2 in Figure 1). In the first step, an annotator can utilize the past example anchors to quickly search for where to place the lower bound of an item using comparisons to work up the scale and finding the position where they can no longer confidently decide that the closest reference should be lower on the scale than the annotated item. Similarly, in the second step, an annotator establishes the upper bound working down from the scale and stopping when they can no longer identify a reference item as higher than the annotated item.

Positions of anchor items on the scale are themselves internally represented by ranges. During each step, the anchors are visualized using the corresponding opposing bound: when finding the *lower* bound for an item, anchor items are placed on the scale according to their *upper* bound values and vice versa for the upper bound (shown in Figure 1). This two-step process allows an annotator to easily establish a range that is intuitive and meaningful—it represents the range where the annotator is no longer able to confidently distinguish items.

3.3 Cold Start Process

Annotation of any item in the Goldilocks process requires there to be previously-annotated items using the same scale in order to populate the grounding examples and comparisons. However, if prior annotations do not exist yet, they must be created in what we call the cold start process.

The cold start process (shown in Figure 3) consists of two steps—representative example selection and placement on a scale. In the example selection step, Goldilocks draws a certain amount of un-annotated examples randomly from the set of data to be annotated. An annotator can then adjust this set by requesting to draw additional random examples or dropping existing examples. The goal is to adjust this set to be more representative such that there are at least a certain number of examples in the set (defined based on task) and that the examples are maximally different from each other. A similar sample and replace approach was used in Alloy [14] to bootstrap good seed sets for clustering. In the placement step, the annotator successively places all the examples onto an

absolute rating scale by comparing them against each other, with the ability to adjust the position of any item on the scale. The scale can be blank at the outset or be initialized with text anchors as shown in Figure 3.

The cold start process can be completed with recruited annotators, where the resulting placements are aggregated across them to create the set of seed examples that become the first set of Goldilocks example anchors. Alternatively, the cold start process can be completed by the task designer or by domain experts, making it a way for requesters to specify a scale without having to design a set of training instances. In this case, the steps in the cold start process are used to assist the exploration of the dataset. Once additional items have been annotated using Goldilocks, the set of anchor examples can be augmented with this newly annotated data. If desired, the initial seed examples can be re-annotated by removing them from the scale and re-introducing them as new items to be placed in an iterative improvement fashion.

3.4 Implementation Details

We outline specific details about our implementation of Goldilocks that we use for experiments. We implemented Goldilocks based on a custom slider component using JavaScript, HTML, and CSS. Global grounding examples were incorporated as part of the scale via fixed anchor tick markers below the scale. Examples were then rendered in a fixed size box attached to each tick mark. Images were scaled to cover the box, and short text was presented as as scrollable content within each box (Fig 2). The interface selects global grounding examples by sorting the set of potential examples and progressively selecting examples that are at least a certain minimum distance from each other. As an annotators scrolls the slider handle, we dynamically search for immediate neighbor examples above and below the slider position and render them as additional anchors placed among the global grounding examples. Neighbor examples are also placed next to the item being annotated to facilitate comparison. Vertical positioning of the rendered anchor examples is dynamically adjusted so that they never visually overlap with each other.

As our experiments were conducted on the Amazon Mechanical Turk crowdsourcing platform, we also implemented a gated training [50] phase for each of the annotation experiments. This phase focuses on training the crowd workers to use the annotation interface rather than annotating a specific task domain, so we used a common training example based on age estimation across all domains. Workers are presented with a prompt describing the task and interface, including specific actions that can be performed using the interface. As workers complete each annotation step for the training task, we check their partial answers against the reference and provide just-in-time feedback if they make a mistake. Once the worker accurately completes the training task, they will progress into the actual annotation task and given the specific instructions for the domain they are annotating. We implemented some basic quality control measures to prevent gaming of the task such as requiring workers to have interacted with the slider before they are allowed to proceed onto the next item.

4 EXPERIMENTS

In order to answer the research questions behind our Goldilocks designs, we conducted annotation experiments using data from 3 domains on the Amazon Mechanical Turk (AMT) platform and using interfaces that isolate specific aspects of Goldilocks for experimentation. Specifically, we tested the following hypotheses:

- RQ1: Does grounding with examples improve consistency?
  - H1-a: Using example-based anchors reduces the amount of disagreement between annotators on ratings of items compared with using semantic text descriptions as anchors.

  - H1-b: Including an annotator's own annotations from the session as additional anchors results in improved self-consistency reflected by less disagreement with their past placement when placing items again.
- RQ2: Does the range-based process create robust output for understanding relationships between items?
  - H2-a: Range annotation captures item resolution and thus can more accurately model distributions of pairwise relationships (more than, less than, indistinguishable) compared to distributions produced by comparing single value annotation output.
  - H2-b: Resolution of items captured using range annotation are better for modeling pairwise relationships than resolution captured through inter-annotator (dis)agreement.
- RQ3: Does the uncertainty about items captured through the size of the ranges correlate with uncertainty captured in the form of inter-annotator disagreement in traditional semantic scale absolute ratings?

4.1 Annotation Task Design

We describe in more detail the task design we used in our annotation experiments, including interfaces derived from Goldilocks and ones from traditional annotation. Unique crowd workers were recruited to use one of the following interfaces to provide annotations for a group of examples:

- **Single Value with Semantic Anchors (SV-SA)**: In each step, annotators are are asked to find a slider position that represents the placement of one item in the annotation sequence using a semantic scale as reference (Figure 4 top).
- **Single Value with Example Anchors (SV-EA)**: In each step, annotators are asked to find a slider position that represents the placement of one item in the annotation sequence using a scale anchored by other example item instances (Figure 4 bottom). Depending on the experiment and condition, the annotator's past placements in earlier steps may become additional anchors for steps in the future.
- **Pairwise**: Annotators were asked to compare all pairs of items. For each step in the annotation sequence, an annotator was presented with 1 reference item and a list of items it has not been compared to yet. For each item, the annotator was asked to judge the relationship of that item compared with the reference item ($>$,$<$,$\approx$).
- **Range with Hybrid Anchors (R-HA)**: This represents the full proposed Goldilocks design. Annotators are given both semantic labels and example instances as reference anchors. For each item, an annotator is first asked to place a lower bound marker for the item followed by placing an upper bound marker. Ranges annotated in earlier steps are incorporated as additional anchors.

Our first study (4.5) examines whether example anchors (**SV-EA**) improve agreement between annotators compated to semantic anchors (**SV-SA**). Following that, our second study (4.6) examines whether including an annotator's past placements improves within annotator consistency when using the **SV-EA** annotation design. Finally, in our last study (4.7), we collect ground truth pairwise relationships directly using the **Pairwise** interface, and compare how well we can recover the distribution of these relationships using data from the traditional single-value semantic anchor approache (with **SV-SA**) with that of the full Goldilocks range annotation design (**R-HA**).

In all cases, annotators were first given a brief gated "interface training" instructional stage where they are guided to annotate a single item (based on an age estimation domain) using the annotation interface they were assigned. Instructions are provided during the process to guide them through using the interface and feedback is given if the annotator makes a mistake in the annotation. Once an annotator completes the annotation process without mistake, they are given

details about the actual task domain they are annotating. Each annotator is then prompted to annotate a sequence of items using the assigned condition's interface.

4.2 Annotation Domains and Datasets

We selected the following 3 annotation domains to to conduct annotation tasks: TOXICITY, SATIETY and AGE. These domains were selected to represent common types of rating tasks that have subjective aspects where a Goldilocks style approach to annotation could be desirable. These tasks also span two different modalities, short text and image, which closely align with rating tasks commonly conducted.

4.2.1 Toxicity. For this task domain, annotators judge the degree of toxicity in a short online comment, estimating how strongly the author of the comment intended to offend. Research has demonstrated that human judgments of online toxicity vary considerably from rater to rater due to subjectivity of the task [57]. The TOXICITY domain represents a short text annotation task where annotators compare pieces of text that only consist of a couple of sentences. Similar tasks include judging fluency of text generation or judging text sentiment. To produce the annotation dataset for this domain, we sampled a 50:50 label-balanced subset of 100 comments from the Jigsaw comment toxicity classification challenge dataset [70] behind the Perspective API[2] which contains Wikipedia comments and binary labels of toxicity. Only comments that had between 4 and 280 characters (after markup removal) were sampled. When presenting the task to crowd workers, we borrow Perspective API's definition of a toxic comment: 'a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion'. We also contrastively define healthy comments as those 'relevant to the discussion' and further note that comments 'can express disagreement'.

4.2.2 Satiety. For this task domain, annotators judge how filling (satiable) is the food depicted in an image, taking into account the type of food and the portion size. The SATIETY domain represents an annotation task that contains uncertainty in the visual modality. Prior research has shown that while pairwise comparisons of food for expected satiety can result in robust ratings, personal familiarity also resulted in biases [11]. We produced the annotation dataset by selecting a subset of food types from the Food-101 dataset [7] and then sampling images for each selected food type up to a total of 80. One round of manual inspection was also done to verify food was clearly discernable in all images.

4.2.3 Age. For this task domain, annotators estimate the age of the subject depicted in a photo. The AGE domain is another annotation task in the visual modality that contains uncertainty, however age is grounded to a concrete scale that we expect most people to be already familiar with. We produced the annotation datset by sampling a subset of 100 portrait images from the FG-NET face dataset [26].

4.3 Anchors for each Domain

To maintain consistency across experiments, we defined a set of text-based semantic differential scale anchors and a set of example anchors for each domain that was held constant across experiments. For the semantic scale anchors, we used text descriptions similar to 7-point Likert or semantic differential scales. Example anchors consisted of 7 roughly evenly spaced in-domain items each associated with a position on the scale.

For the TOXICITY domain, we used the following text descriptions for semantic scale levels: "1 - Not Toxic at All", "4 - Somewhat Toxic" and "7 - Extremely Toxic". Other levels (2, 3, 5, 6) on the scale were presented as a number without any associated description. The 7 example anchors

[2] https://www.perspectiveapi.com

were manually picked from a set of annotated examples produced from a pilot run of the cold start process with crowd annotators.

For the SATIETY domain, we used the following text descriptions for semantic scale levels: "1 - Very Hungry', "2 - Somewhat Hungry", "3 - Almost Satisfied", "4 - Satisfied', "5 - Full", "6 - Very Full", and "7 - Can't Finish". The 7 example anchors were produced by the authors producing gold annotations directly using the cold start process interface.

For the AGE domain, we used text scale levels based on numeric age values ranging from "0" to "60+" incrementing in steps of 10. The 7 example anchors were picked by finding all images corresponding to each semantic age level and then drawing a random one at each level and assigning its value to be the ground truth age.

4.4 Crowd Annotator Recruitment and Compensation

We recruited annotators for our experiments from the Amazon Mechanical Turk crowdsourcing platform from the United States with the qualification of approval rate no lower than 90% and over 1000 approved HITs completed in the past. Across all studies, annotators were only allowed to participate in annotation if they had both not used the corresponding interface and not annotated the domain before. Overall, we recruited 655 unique workers across all 3 studies with an additional 44 unique workers who only participated in the pairwise annotation used to establish the ground truth for Study 3. For all annotation tasks, we set a base pay of $0.10 which was given if the worker completed the training phase. Remaining compensation was distributed in the form of a bonus based on the interface being used and the number of items annotated.

Participants assigned to the **Single Value** tasks (both with **Semantic** and **Example** anchors) were given a per-item bonus of $0.03 (for annotating a group of 10 or 20 items). Participants assigned to the **Range** tasks were given a per-item bonus of $0.05 (a total of 10 items). Participants assigned to the **Pairwise** annotation tasks were given a per-comparison bonus of $0.01 (a total of 45 comparisons). We set pay based on our estimate of time needed taken from pilot studies and used completion bonuses to correct for any discrepancies. Based on condition, a final completion bonus of $1.00, $0.50, or $1.00 for each of the previously mentioned interfaces respectively was provided. We distributed the final bonus in 2 batches as the initial completion bonus values we set for the tasks resulted in a measured hourly pay that was lower than desired. The final hourly rate measured between $9.70 and $10.90 across the various domains and interfaces when assuming the median work time for each interface.

Manual quality checks were conducted on cases with a large number of similarly annotated values across different items (e.g., consistently placing at 0 or 1) as well as abnormally short work time, resulting in removal of 5 workers (and re-collection of corresponding annotations) across all experiments. Removed workers were included in the counts of recruited workers above. Within the removed workers, those intentionally spamming across their entire sequence of annotations (choosing the exact same placement for all items) only received the base pay for the task.

4.5 Study 1: Evaluating Consistency Between Annotators

We first explore whether example-based grounding presented in Goldilocks can improve consistency between different annotators (H1-a). For this experiment, we assigned each annotator to one of two conditions: SEMANTIC, where they were given 7-point text-based semantic anchors and presented with the **SV-SA** interface; or EXAMPLE, where they are given 7 example instances placed onto the scale using the **SV-EA** interface. For each domain, the anchors used are detailed in 4.3. We drew example anchor instances for the TOXICITY and SATIETY domains from past pilots of semantic differential scale annotation on a disjoint set of items, using average rating to establish their initial

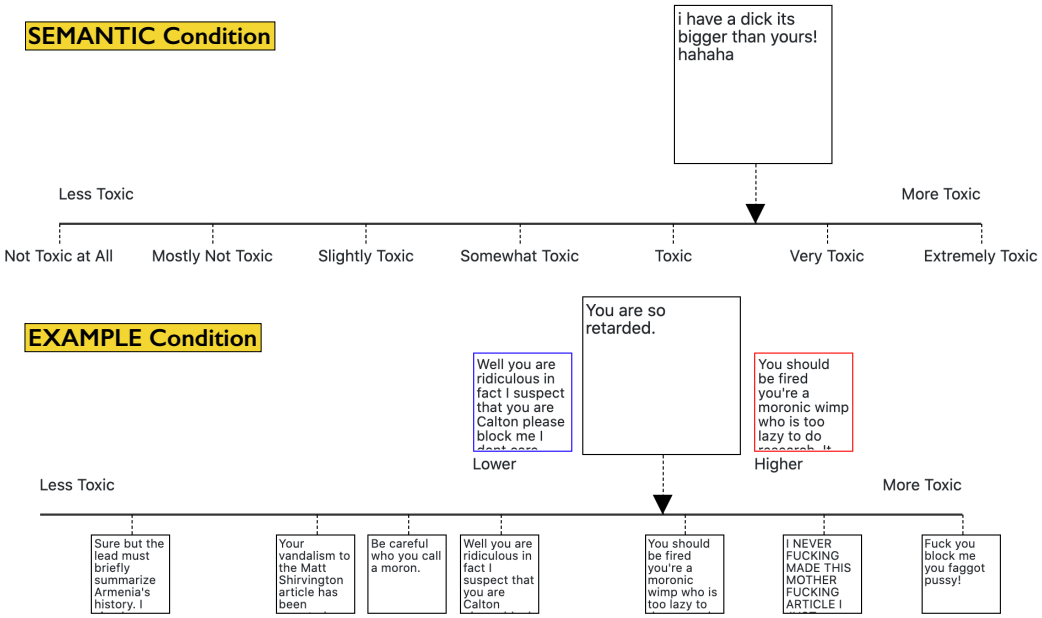


Fig. 4. Screenshot showing the two interfaces conditions (top: SEMANTIC, and bottom:EXAMPLE) used to evaluate consistency consistency between annotators. Examples shown in figure are from the toxicity domain pilot tasks. (Content warning: toxic comments including offensive and swear words are shown in their original form as a part of this figure.)

placement. For the AGE domain, example instances were selected from a separate set of images drawn from the same dataset using the included ground truth age labels for initial placement.

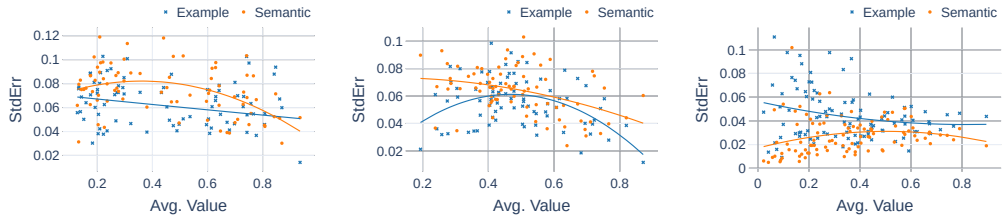
After the training, each annotator was tasked with annotating a sequence of 10 items using the interface of the condition they were assigned. To create sequences, each domain's dataset was shuffled once and then partitioned into equal-sized disjoint sets. Each sequence for each domain was annotated by 10 workers in each of the two conditions. Annotators' placements of items on the scale was mapped as a continuous numeric value within the range [0, 1]. For the TOXICITY domain, the first and last items in each sequence were set to the same item to pilot measurement of within-annotator consistency, so only the 8 remaining annotations were used for analysis in this experiment.

4.5.1 Results. To evaluate the amount of consistency between annotators for each annotated data point, we computed the standard error across annotators as a proxy for the amount of disagreement. We note that the standard error values are comparable across conditions as the range of values on the scale and number of annotators was fixed between all conditions. We also evaluated the significance of any difference by conducting a two-tailed paired t-test on the standard error of each annotated item across each pair of conditions (SEMANTIC versus EXAMPLE) in each domain. A summary of the results are shown in Table 1.

We observed a statistically significant decrease in value disagreement across annotators for the TOXICITY and SATIETY domains, providing support for hypothesis H1-a. However, we observed a statistically significant increase in disagreement across annotators for the AGE domain, which contradicts H1-a. We then plotted the disagreement (standard error) in both conditions for each

Table 1. Results for the experiment measuring consistency between annotators comparing between SEMANTIC and EXAMPLE conditions. Average disagreement is calculated as the standard error (over 10 annotators) for each instance averaged across all annotated instances. Significance testing done as a paired t-test across conditions for disagreement. We also examine how much of the 0-1 scale is being used by annotators on average in each condition by averaging each annotator's minimum and maximum rating values.

Domain	Condition	Avg. Disagreement	Significance	Scale Util. ($\overline{\text{Min}}$, $\overline{\text{Max}}$)
TOXICITY	SEMANTIC	0.07348	$P < 0.001$	0.773 (0.103, 0.876)
	EXAMPLE	**0.06379**	Very Significant	0.794 (0.104, 0.899)
SATIETY	SEMANTIC	0.06373	$P < 0.005$	0.603 (0.230, 0.833)
	EXAMPLE	**0.05548**	Significant	0.635 (0.166, 0.801)
AGE	SEMANTIC	**0.02765**	$P < 0.001$	0.696 (0.054, 0.751)
	EXAMPLE	0.04443	Very Significant	0.593 (0.072, 0.665)



(a) TOXICITY domain. (b) SATIETY domain. (c) AGE domain.

Fig. 5. Scatter plots of disagreement between workers (as measured by standard error) for each item plotted against the mean annotated value of each item. Trendlines represent a fit with a degree 2 polynomial.

item against the mean value across both conditions in each domain to understand the behavioral differences we see with the age domain as shown in Figure 5.

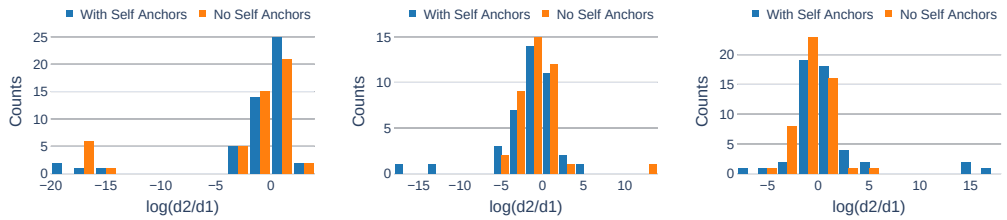
We find that the pattern for disagreement in the SEMANTIC condition is consistent with behavior observed in prior work [69] for similar domains with subjectivity and uncertainty. However, we note that overall disagreement between annotators was lower in the AGE domain compared to the other two domains. We also noted that scale utilization was similar in both conditions for the TOXICITY and SATIETY domains, exhibiting a slightly increase in utilization of the full scale in the EXAMPLE condition. Prior work in psychology has shown that increased spacing of items has relatively minimal effect on accurate placement when items are discriminable [62] so we don't expect this slight increase in scale utilization to affect disagreement levels. However, opposite to the other domains, the utilization of the scale in the AGE domain was 10% *lower* for the EXAMPLE condition. We hypothesize that unlike the TOXICITY and SATIETY domains, estimating age from appearance is a domain where a numeric age scale is actually more consistently understood by human annotators, thus example anchors provide no further benefit to annotators in understanding the scale. The scatter plots in Figure 5c indicate that uncertainty for younger subjects was much higher in the EXAMPLE condition. Combined with the lower scale utilization we observed for EXAMPLE, we hypothesize that uncertainty about judging exact age is higher for older subjects. As we only show example-based anchors in the EXAMPLE condition, this increased uncertainty about the reference images depicting older subjects may have resulted in more hesitation to use the higher

values on the scale. This suggests that: (1) comparisons with anchor examples mostly benefit cases where shared understanding of the scale is low, and (2) example-based anchoring should be used in *addition* to semantic anchors as *only* using example anchors can be detrimental to consistency if the domain is one where the semantic scale has a high degree of shared understanding already. Drawing from this experiment, our full Goldilocks annotation process uses both example-based anchors and semantic anchors to frame the scale.

4.6 Study 2: Evaluating Consistency Over Time Within Annotator

For our second experiment, we explored the effect on self-consistency resulting from including an annotator's own past annotations as additional reference examples augmenting an initial seed set (H1-b). The example-based **SV-EA** interface was used for this experiment, with each annotator was assigned one of the two conditions: CONTROL, where only the seed set examples was used for reference anchors; or AUGMENT, where an annotator's own past annotations in the same session were included along the seed examples as references. Since we are interested mainly in the effect on self-consistency, we reduced the initial set of seed examples to just 3 examples for each domain drawn as a subset of the 7 example instances used in the EXAMPLE condition of the previous experiment. We took the items corresponding to the lowest, highest, and median ratings.

The items in each domain were shuffled and then partitioned into sequences of size 20, resulting in 5 sequences for the TOXICITY and AGE domains and 4 sequences for the SATIETY domain. Each annotator was given interface training and then subsequently tasked with annotating one of the sequences (of 20 items). To probe for changes in the rating of an item, we replaced the 10th and 20th items in each sequence above with repeats of the first item, which we will refer to as the probe item. When the probe item is annotated in the AUGMENT condition, the annotator's own past annotation for the probe item will be withheld from the set of reference items. We measure Δ_1 as the size of the value change between the first and second annotation attempts of the probe item and Δ_2 as the size of the value change between the second and third annotation attempts of the probe item.



(a) TOXICITY domain. (b) SATIETY domain. (c) AGE domain.

Fig. 6. Histogram of distance ratios between first re-annotation and second re-annotation of the probe item on a log scale. Negative values indicate more decrease in disagreement with the annotator's own answers while positive values indicate more increase in disagreement with the annotator's own answers. Ratios were smoothed using Laplace smoothing with $\epsilon = 10^{-8}$.

4.6.1 Results. From Table 2 we can see that for most domain condition pairs, the absolute amount of an annotator's disageement with their past rating tends to exhibit a natural decrease as they get familiarized with the scale. Since the magnitude of initial self-disagreement for the probe item varies for each annotator, comparing absolute change in self-disagreement can be misleading as the same *proportional* change in self-disagreement will reflect as a larger *absolute* change. To

Table 2. Table breakdown of the change in rating for the probe item (compared to its last most recent rating) when re-annotated for the first time (Δ_1) and when re-annotated the second time (Δ_2). The "Top Avg. Δ" columns represent the averages when only considering the instances where Δ_1 was among the top 30% most uncertain.

Domain	Condition	Avg. Δ_1	Avg. Δ_2	Top Avg. Δ_1	Top Avg. Δ_2
TOXICITY	No Self (CONTROL)	0.105	0.062	0.244	0.077
	With Self (AUGMENT)	0.133	0.090	0.347	0.095
SATIETY	No Self (CONTROL)	0.140	0.086	0.308	0.171
	With Self (AUGMENT)	0.126	0.052	0.286	0.027
AGE	No Self (CONTROL)	0.110	0.063	0.280	0.133
	With Self (AUGMENT)	0.063	0.066	0.157	0.067

account for these factors, we instead look to the self-disagreement *ratio* (Δ_2/Δ_1) as a measurement for the proportional decrease (or increase) in self-disagreement. Ratios below 1 indicate that self-disagreement has decreased while those above 1 indicate an increase. In Figure 6, we show a histogram of this ratio on a log-scale for each condition in this study.

Our first step is understanding whether self consistency improves over time simply from doing the task and being exposed to more examples. We conducted a sign test for each of the task domains and find that in the TOXICITY domain, self consistency does improve over time ($P < 0.005$) for both CONTROL and AUGMENT conditions. Self consistency was not found to have a significant across-the-board improvement in any of the other domains. Comparing across the two conditions, we did not measure significant effect on self-disagreement ratio in any of the 3 domains.

We then hypothesized that effect on self-consistency may not be uniform across all probe items—if an annotator already has low self-disagreement in the first re-annotation round (Δ_1), it likely implies there is little uncertainty about the placement of the item and thus we shouldn't expect further improvements. Considering this, we now look at only the top 30% 'most uncertain' annotation sessions for each domain and condition combination, as sorted by decreasing Δ_1. This set consists of 15 sessions for the TOXICITY and AGE domains and 12 for the SATIETY domain. In this high-disagreement subset of sessions, we find that augmenting reference examples (AUGMENT) with past annotations in the session does result in a larger proportional reduction in self-disagreement (reflected through self-disagreement ratios) when compared to CONTROL for both the TOXICITY and SATIETY domains. For the SATIETY domain, median proportional decrease in self-disagreement was 0.076 (92% reduction in self disagreement) for the AUGMENT condition compared to 0.263 (74% reduction) for the CONTROL. The median ratios were 0.190 (81% reduction) and 0.310 (69% reduction) respectively for the TOXICITY domain. However, the limited amount of data points in these groups means we do not have statistical power to claim significance. Overall, we don't find sufficient support for H1-b, but we note a pattern of improvement in self-consistency for items with high initial self-disagreement when including an annotator's own prior annotations as additional references. Similar to the previous section, we were unable to observe benefit of augmenting reference examples on the AGE domain, likely due to the already limited utility of reference examples in this domain.

4.7 Study 3: Evaluating Range Annotation

For the final experiment, we explored how robustly ranges produced by the two-step annotation process in Goldilocks reflect properties of relationships between items. In this experiment, annotators were asked to annotate a sequence of items using the full Goldilocks two-step annotation process (using the **R-HA** design shown in Figure 1). The annotation experiments were conducted on the TOXICITY and SATIETY domains with sequences generated by shuffling each dataset and partitioning the dataset into groups of size 10, resulting in 10 and 8 groups respectively for the two domains. We then recruited 5 annotators to annotate each sequence in each of the domains.

At the start of the task, each annotator was first trained on how to use the two-step annotation system described earlier in Section 3.2 by annotating a sample task with guidance given during each step. After the annotator completes the training example item, they then proceed to annotate the assigned sequence of 10 task items. To seed the initial reference examples, we used the same reference anchors as used in the first experiment. We also included each annotator's own annotations as anchors during their annotation in a similar way as the AUGMENT condition in the second experiment.

4.7.1 Establishing Pairwise Relationship Distributions. In order to measure ground truth distributions over the pairwise relationships, we recruited separate workers and used the **Pairwise** design to directly collect pairwise judgments on relationships (>,<,≈) between all pairs of items in each group. Distributions across the 3 relationship types were then created by counting the proportion of annotators indicating each type of relationship across for each pair of items. These distributions reflect the degree of disagreement among annotators for the pairwise relationship.

We then considered how one would recover similar distributions across relationships for pairs of items using the traditional approach of single-value absolute rating scales based on semantic anchors, creating two alternative baselines. Since the traditional approach cannot simultaneously elicit item ambiguity and agreement, producing a similar distribution would involve a tradeoff.

For the **Direct** baseline, we assume that there is no item-level ambiguity, meaning that even local pairwise comparisons can be made by directly comparing the raw values from the absolute rating. For example, we count an annotator as indicating a "$>$" relationship on a pair (a, b) if their single rating scores indicate $r_a > r_b$. One can generally expect this to be reliable when a and b are far apart on the scale but it can be much less reliable for close neighbors.

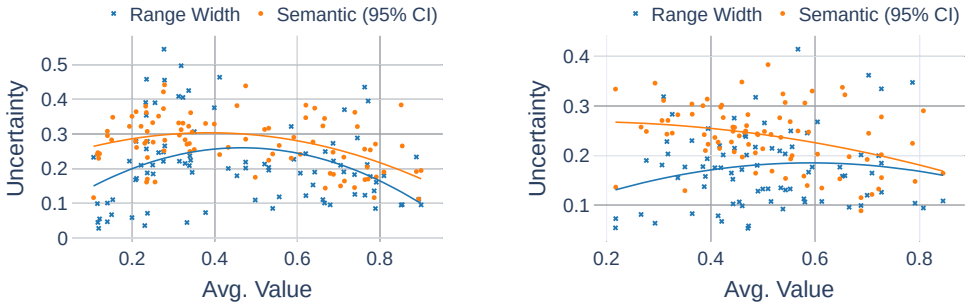
For the **Infer** baseline, we assume that all disagreement observed between annotators reflects the ambiguity of the item. In this case, we aggregate the individual ratings into a *single* 95% confidence interval for each item by measuring the mean and standard error between these samples. We then infer the relationship between of a pair of items by comparing the confidence intervals, treating overlapping intervals as indicating a relationship of 'indistinguishable (≈)'. In this case, the distribution across relationships for a pair would see all the probability mass allocated to the single relationship measurement produced by the comparison.

Finally, with Goldilocks annotation, we have range evaluations on a per-annotator granularity. For each annotator, we can use their range labels to find the relationship between two items, treating overlaps as indicating ≈. We can then produce a distribution by counting the proportion of annotators indicating each relationship. With Goldilocks we don't need to make tradeoffs between measuring item ambiguity and agreement.

4.7.2 Results: Recovering relationships between items. To compare and quantify how robustly each of these methods recovers relationships, we measured the Wasserstein distance between relationship distributions for each of the 3 approaches in 4.7.1 and the ground truth relationship distributions collected through pairwise comparative rating. Table 3 shows that among the 3

Table 3. Comparing the quality of the pairwise relationship distributions as recovered by (1) **range**s collected in Goldilocks, (2) **direct**ly comparing raw values picked by each annotator, and (3) indirectly using ranges **infer**red from the 95% confidence intervals. Details in 4.7.1. Wasserstein distance to the ground truth distribution (collected directly using pairwise comparisons) was computed for each case. Goldilocks ranges produce distributions the closest (least distance) to the ground truth.

Domain	Avg. WD (Range)	Avg. WD (Direct)	Avg. WD (Infer)
TOXICITY	**0.332314**	0.366944	0.450556
SATIETY	**0.424352**	0.449444	0.597222



(a) TOXICITY domain.

(b) SATIETY domain.

Fig. 7. Comparison of uncertainty measured as range sizes from Goldilocks annotation with uncertainty measured through standard error confidence intervals from traditional single-value semantic scale annotation. Trendlines represent a fit with degree 2 polynomials.

methods to produce distributions over pairwise relationships, recovering distributions using range labels most accurately agrees with the ground truth distribution, supporting H2-a.

We found that using inter-annotator agreement to infer the inherent ambiguity (referred to by prior works as "resolution") of items results in an over-estimate of the amount of ambiguity. In the TOXICITY domain, 43.5% of the relationships that were distinguished in the ground truth distribution collected directly through pairwise comparisons were inferred to be "indistinguishable", with this ratio as high as 68.1% in the SATIETY domain. In contrast, ranges over-estimate ambiguity (under-estimating resolution) only about half as often, with 22.1% and 30.9% respectively. This supports the idea that ranges are a better model of resolution (H2-b).

4.7.3 Results: Comparing aggregation uncertainty with range sizes. Finally, we explored differences in the type of uncertainty measured through Goldilocks annotation ranges sizes with uncertainty measured by confidence intervals of annotations using semantic scales. We hypothesize that since ranges focus on capturing resolution (distinguishability against peers) of items, the resulting uncertainty represented by the size of ranges will be different than uncertainty represented by inter-annotator disagreement metrics, though the two may still be related.

First we look at the behavior of the two kinds of uncertainty measurements across the range of values on the scale. Figure 7 plots the two kinds of uncertainty: average size of ranges and 95% confidence intervals for semantic scale annotation values. We find that overall range sizes represent uncertainty lower than that measured by 95% confidence intervals from aggregating semantic scale

annotation ($P < 0.001$). This makes intuitive sense as we would expect item level resolution to be a tighter uncertainty. We also find that in the TOXICITY domain, both types of uncertainty behave similarly with respect to extreme values on the scale corresponding to lower values of uncertainty in definitions. In the SATIETY domain, however, we found that lower values (corresponding to foods depictions that are less satiating) corresponded to larger uncertainty in the form of disagreement but not with range sizes. We think this may result from higher disagreement about what foods are not satiating among different annotators but with annotators each confident about their own determination of satiety (high resolution/distinguishablity of items).

Looking at correlation between the values produced by the two types of uncertainty, we observe only very weak correlation between range sizes and confidence intervals (scaled standard error) for both TOXICITY and SATIETY domains with $R^2 < 0.01$ in both domains. This indicates that the uncertainty we measure with ranges does not have significant correlation with inter-annotator disagreement measures like standard error (RQ3). We note that with range annotations, inter-annotator disagreement measures can be further computed for the range bounds themselves to evaluate disagreement separately from item uncertainty (resolution) captured by ranges. However, as single-value semantic scalar annotations can't facilitate separation of the two uncertainty types, we are unable to make direct comparisons.

5 DISCUSSION

In the prior sections, we demonstrate that the ideas of grounding absolute rating scales with examples and explicitly capturing item-level measurement resolution can be beneficial for more consistent and robust annotation of subjective domains lacking shared understanding of absolute ratings scales. In this section, we will discuss some of the other considerations in adapting Goldilocks as a full annotation technique, including examining the annotation efficiency (in terms of work time) of Goldilocks compared to hybrid application of traditional methods and envisioning how Goldilocks may be scaled up to multiple annotation sessions using iterative-improvement processes. We will also discuss limitations of the Goldilocks process and potential avenues for future work.

5.1 Annotation Efficiency and Cost of Range Annotations

One of the main advantages of the Goldilocks annotation process is the ability to capture item-level ambiguity and disagreement between annotators simultaneously through the use of range annotations. However, separating these sources of uncertainty comes at an extra cost for the data collection process—even though range bounds in Goldilocks can be collected with low overhead compared to traditional absolute rating, the tasks can be more work for the requester to set up. This presents a tradeoff for practitioners when deciding whether the higher quality of data is worth the cost. Prior work simulating data annotation tasks inspired by measuring objective properties has shown that, given a fixed budget, some learning algorithms actually benefit more from a larger amount of lower-quality annotations on novel examples rather than higher-quality annotations on fewer items [47]. Indeed, for these tasks where disagreement is likely caused by noisy perception, it's likely that a practitioner will see relatively little benefit by separating item-level ambiguity from annotator disagreement. However, with the rising demand for training data in domains that involves subjectivity or nuance, understanding and accounting for sources of uncertainty and limitations within the data itself has become increasingly important for building models that are *trustworthy* rather than just more *performant* [5]. Separating disagreement from inherent ambiguity using range-based annotation can also offer better transparency about the annotation process and data produced, allowing for the potential to diagnose model limitations and human biases even into the future. In these cases, the higher cost of setting up Goldilocks annotations can be justified by the richer information that can be derived from range-based rating data.

Of course, Goldilocks is not the only approach to capture both item-level ambiguity and disagreement. It is possible to use traditional absolute and comparative rating to separately collect scalar annotations and pairwise comparisons to recreate absolute rating estimates and pairwise relationship distributions. We also wanted to understand whether Goldilocks can provide efficiency benefits when compared to hypothetical hybrid approches using only a combination of traditional annotation interfaces. We look at the work time taken by crowd workers in our various experiments to extrapolate the effort necessary for such an approach. Assuming a task group size of 10 items, we find that the Goldilocks two-step workflow results in a median work time (including both training and annotation) of 429.5s per worker per task group on the SATIETY domain and 592.5s per worker per task group on the TOXICITY domain. Collecting only single value rating annotations with Likert-style anchors takes a median work time of 307.5s per worker per task group on the SATIETY domain and 238s per worker per task group in the TOXICITY domain. Finally, comparative rating on a group of size 10 implies 45 pairwise comparisons to capture full pairwise relationships, which takes a median time of 513.5s per worker per group and 502s per worker per group for the two domains respectively. Thus we expect that at the same level of redundancy for annotations, Goldilocks can be 20-48% more efficient through the use of our two-step range-based annotation that collects ratings and relationship distributions together. Consistency improvements of Goldilocks may be able to push efficiency further in practice by requiring a lower amount of redundancy to achieve the same level of agreement.

5.2 Goldilocks and Iterative Improvement

So far in this paper we have examined the ideas presented in Goldilocks only for single annotation sessions where we didn't need to update the anchor examples beyond incorporating an annotator's own ratings. In order to scale up to larger datasets, it becomes necessary to perform annotations over multiple sessions which involve using aggregation approaches to iteratively construct an updated set of anchors. To achieve this we envision a process based on the idea of iterative improvement [49].

In each round of iteration, a group of annotators individually annotate a subset of the dataset, sharing a 'seed' set of anchor examples used to ground the interpretation of the scale, with their own annotations also incorporated as they progress along the annotation session. Once all annotators have completed the session, the annotations collected will be aggregated into a new set of seed examples used to ground the next round of iteration. In addition to progressively annotating new examples, this iterative process may also be used to revise past annotations, such as those created during the cold start process. This can be accomplished by first removing the items to be revised from the set of grounding examples and then re-annotating them as new items in an iteration. This process of periodically aggregating annotations and then re-seeding anchor examples can serve as a method to scale up annotations while ensuring a stable scale as annotators place items.

We believe that this represents a feasible design for scaling up annotation, and we envision further work can be done to explore options for aggregation and re-annotation strategies as well as evaluate their effectiveness. We also see potential for using iterative improvement as way to dynamically re-calibrate the definition of scales to account for distributional shifts over time. For example, a scale that can dynamically adapt to improving quality of machine summarization systems can be adapted as a living benchmark. We think the ideas presented in Goldilocks for single annotation sessions provide a first step into building an effective iterative workflow.

5.3 Limitations and Future Work

While Goldilocks provides a path to more consistent scalar annotation that also captures uncertainty, we also recognize that the current design is still subject to some limitations which we believe can be good avenues for future work.

5.3.1 Creating High Quality Seeds in Cold Start. The cold start process in Goldilocks provides a way to generate the initial seed set of grounding examples that enable the comparisons and consistency benefits of Goldilocks. However, the quality of this initial set of seed examples can also influence whether consistency benefits can be realized. We observed some of these limitations when experimenting with example-based anchors in the AGE domain. A good seed set should consist of examples that achieve both good coverage of the scale and have low ambiguity themselves. When the seed set achieves good coverage over the scale, the comparative process can allow seed examples that are distinguishable to quickly be excluded from the range of the annotated item, resulting in measurement resolution that mainly depends on the number of examples in the seed set. However, a set of examples that is not representative of the full range of items to be ranked can lead to issues of scale drift when these examples (that annotators may desire to rate higher or lower than the current implied bounds of the scale) are encountered in the future. The current cold start process provides some mitigation to the issue of representativeness by incorporating a 'resampling and replace' phase to increase the diversity of items in the seed set. However, for sufficiently large datasets this may not be enough to capture rare items that are also outliers for the scale. For future work, we envision enabling the ability for annotators to rescale the visible scale itself through an interaction similar to zooming in or out, allowing the annotation of items that lie outside the current extremes of the scale when they are encountered.

Another current limitation of the cold start process is that the cold start design cannot effectively capture item ambiguity as we only elicit a single label for each reference item. In pilot studies we found it infeasible to introduce ranges into the cold start process as there are no anchors to compare against to effectively determine these ranges. It is possible to have suboptimal seed sets where the seed items can have high ambiguity themselves, thus acting as a lower bound on range sizes. We hypothesize that the iterative improvement process in 5.2 may offer a way to limit the impact of the cold start seeds if we can conduct subsequent annotation rounds where we can instead seed with regular annotated range data, though we leave exploration of this to a future study.

5.3.2 Addressing Long-form Tasks and Context. Some common tasks where crowd scalar ratings are desirable, such as evaluating relevance, conciseness, fluency, or faithfulness of summaries produced by text summarization models, can depend on understanding long-form context (e.g., a news article) or even multiple documents [25]. While we have shown that Goldilocks can support annotation domains based on small amounts of text (1-2 sentences) using a similar interface as the one used for images, long-form text will require a different design for conducting comparisons both with the global scale and local neighborhood.

Additionally, interactions in Goldilocks assume that items can be compared against other items in the same dataset. However, when rating items with context, such as summarization or translation, it is likely that reasonable comparisons can only be made with certain other items sharing the same context (i.e., alternate summaries/translations of the same source). A potential avenue for future work extending Goldilocks may exist in introducing virtual views to the Goldilocks scale that enable contextual comparisons on the scale by only exposing items sharing the same context. Future work on an algorithm for determining optimal global example anchors could also take into account aspects that could make comparison easier, such as similarity to the item being annotated.

5.3.3 Working with Density. One of the strengths of Goldilocks is the ability to use past annotations from any source, including data from existing datasets to establish grounding for a scale. By providing past annotations from a dataset as reference examples, it will be possible to augment the dataset in a way that is consistent with past examples but also doesn't require building complex rubrics. However, as the set of past annotations increases, it poses potential problems for the local comparison aspect of the Goldilocks annotation process. There are practical limitations on how fine

adjustments can be on a slider-based scale, so as regions on the scale become densely populated by examples, it becomes harder to use local comparisons to find precise upper and lower bounds in those regions. Even small adjustments in a dense region can mean moving across many reference points.

One potential solution to the density problem could come from allowing the scale to be itself scaled, similar to that proposed in 5.3.1. Initially the full view of the scale is presented along with global anchors for coarse navigation. As an annotator narrows down on a dense region, they can increase the zoom level of the annotation scale to span just the dense region across the entire width of the scale, increasing the amount of space and in turn reducing interaction issues caused by density. New global anchors can be selected to allow for quick navigation at the new zoom level.

6 CONCLUSION

In this paper, we present and evaluate Goldilocks, a novel technique to elicit scalar annotations using the crowd that improves on consistency and captures pairwise relationships more robustly. We show that by prior examples can be used as anchors to ground otherwise abstract absolute rating scales (such as semantic or Likert scales) leading to more consistent interpretation between workers. We find that including an annotator's past annotations in a session can lead to more self consistency on items that have high initial uncertainty. Finally, we show that introducing range annotation into absolute rating can enable simultaneous elicitation of both perceived ambiguity on a per-annotator scale while also capturing inter-annotator disagreement. This simultaneous measurement enables a better recovery of pairwise relationship distributions.

REFERENCES

[1] Abhaya Agarwal and A. Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. In *WMT@ACL*.

[2] Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor, and Julia Hirschberg. 2020. A novel methodology for developing automatic harassment classifiers for Twitter. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 7–15.

[3] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*. 1100–1105.

[4] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24.

[5] Umang Bhatt, Y. Zhang, J. Antorán, Q. Liao, P. Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, R. Krishnan, Jason Stanley, O. Tickoo, L. Nachman, R. Chunara, Adrian Weller, and Alice Xiang. 2020. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. *ArXiv* abs/2011.07586 (2020).

[6] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2020).

[7] Lukas Bossard, M. Guillaumin, and L. Gool. 2014. Food-101 - Mining Discriminative Components with Random Forests. In *ECCV*.

[8] C. Bouveyron and S. Girard. 2009. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognit.* 42 (2009), 2649–2658.

[9] Jonathan Bragg, Mausam, and Daniel S. Weld. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (2018).

[10] G. Brown, I. Neath, and N. Chater. 2007. A temporal ratio model of memory. *Psychological review* 114 3 (2007), 539–76.

[11] Jeffrey M Brunstrom, Nicholas G Shakeshaft, and Nicholas E Scott-Samuel. 2008. Measuring 'expected satiety'in a range of common foods using a method of constant stimuli. *Appetite* 51, 3 (2008), 604–614.

[12] Chris Callison-Burch, M. Osborne, and Philipp Koehn. 2006. Re-evaluation the Role of Bleu in Machine Translation Research. In *EACL*.

[13] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[14] Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. 2016. Alloy: Clustering with crowds and computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3180–3191.
[15] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[16] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
[17] Andrew P Clark, Kate L. Howard, A. Woods, I. Penton-Voak, and Christof Neumann. 2018. Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PLoS ONE* 13 (2018).
[18] A. P. Dawid and A. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics* 28 (1979), 20–28.
[19] Michael J. Denkowski and A. Lavie. 2010. Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks.
[20] Shayan Doroudi, Ece Kamar, Emma Brunskill, and E. Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016).
[21] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. 2016. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy.. In *Hcomp*. 32–41.
[22] A. Dumitrache. 2015. Crowdsourcing Disagreement for Collecting Semantic Annotation. In *ESWC*.
[23] A. Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing Ambiguity in Crowdsourcing Frame Disambiguation. In *HCOMP*.
[24] A. Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8 (2018), 1 – 20.
[25] A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, R. Socher, and Dragomir Radev. 2020. SummEval: Re-evaluating Summarization Evaluation. *ArXiv* abs/2007.12626 (2020).
[26] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Yuan Yao, and Shaogang Gong. 2014. Interestingness Prediction by Robust Learning to Rank. In *ECCV*.
[27] Henry J. Gardner and M. Martin. 2007. Analyzing Ordinal Scales in Studies of Virtual Environments: Likert or Lump It! *PRESENCE: Teleoperators and Virtual Environments* 16 (2007), 439–446.
[28] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *EMNLP*.
[29] Mor Geva, Y. Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. *ArXiv* abs/1908.07898 (2019).
[30] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *CoRR* abs/1412.6572 (2015).
[31] Mitchell Gordon, Kaitlin Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.
[32] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, Sofia, Bulgaria, 33–41. https://www.aclweb.org/anthology/W13-2305
[33] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. *ArXiv* abs/1706.04599 (2017).
[34] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3511–3522.
[35] Jordan S. Huffaker, Jonathan K. Kummerfeld, Walter S. Lasecki, and M. Ackerman. 2020. Crowdsourced Detection of Emotionally Manipulative Language. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
[36] Scott Huffman. 2008. Search evaluation at Google. https://googleblog.blogspot.com/2008/09/search-evaluation-at-google.html.
[37] E. Hullermeier and W. Waegeman. 2019. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *arXiv: Learning* (2019).
[38] Tao Jin, Pan Xu, Quanquan Gu, and F. Farnoud. 2020. Rank Aggregation via Heterogeneous Thurstone Preference Models. In *AAAI*.
[39] David Jurgens. 2013. Embracing Ambiguity: A Comparison of Annotation Methodologies for Crowdsourcing Word Sense Labels. In *HLT-NAACL*.

[40] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1637–1648.

[41] Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561* (2021).

[42] Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 811–817. https://doi.org/10.18653/v1/N16-1095

[43] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3075–3084.

[44] Samuel Läubli, Rico Sennrich, and M. Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *ArXiv* abs/1808.07048 (2018).

[45] Weixin Liang, J. Zou, and Zhou Yu. 2020. Beyond User Self-Reported Likert Scale Ratings: A Comparison Model for Automatic Dialog Evaluation. In *ACL*.

[46] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[47] C. H. Lin, Mausam, and Daniel S. Weld. 2014. To Re(label), or Not To Re(label). In *HCOMP*.

[48] Christopher H. Lin, Mausam, and Daniel S. Weld. 2016. Re-Active Learning: Active Learning with Relabeling. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, Arizona) *(AAAI'16)*. AAAI Press, 1845–1852.

[49] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. Exploring Iterative and Parallel Human Computation Processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (Washington DC) *(HCOMP '10)*. Association for Computing Machinery, New York, NY, USA, 68–76. https://doi.org/10.1145/1837885.1837907

[50] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H. Lin, Xiao Ling, and Daniel S. Weld. 2016. Effective Crowd Annotation for Relation Extraction. In *Proceedings of NAACL and HLT 2016*.

[51] Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9.

[52] Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 630–638.

[53] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. Number 47. University of Illinois press.

[54] Y. Qin, Xuezhi Wang, Alex Beutel, and Ed Huai hsin Chi. 2020. Improving Uncertainty Estimates through the Relationship with Adversarial Robustness. *ArXiv* abs/2006.16375 (2020).

[55] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. *ArXiv* abs/1806.03822 (2018).

[56] Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient Online Scalar Annotation with Bounded Support. *ArXiv* abs/1806.01170 (2018).

[57] Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung, and Bernard J Jansen. 2018. Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 88–94.

[58] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[59] João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, L. Ungar, and Chris Callison-Burch. 2019. ChatEval: A Tool for Chatbot Evaluation. In *NAACL-HLT*.

[60] Edwin Simpson and Iryna Gurevych. 2018. Finding Convincing Arguments Using Scalable Bayesian Preference Learning. *Transactions of the Association for Computational Linguistics* 6 (2018), 357–371.

[61] Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 Shared Task on Machine Translation Robustness. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online, 76–91. https://www.aclweb.org/anthology/2020.wmt-1.4

[62] N. Stewart, G. Brown, and N. Chater. 2005. Absolute identification by relative judgment. *Psychological review* 112 4 (2005), 881–911.

[63] N. Stewart, N. Chater, and G. Brown. 2006. Decision by sampling.

[64] Louis L Thurstone. 1927. A law of comparative judgment. *Psychological review* 34, 4 (1927), 273.
[65] T. Tian, J. Zhu, and Y. Qiaoben. 2019. Max-Margin Majority Voting for Learning from Crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 10 (2019), 2480–2494. https://doi.org/10.1109/TPAMI.2018.2860987
[66] C. Völker, T. Bisitz, R. Huber, B. Kollmeier, and Stephan M. A. Ernst. 2018. Modifications of the MUlti stimulus test with Hidden Reference and Anchor (MUSHRA) for use in audiology. *International Journal of Audiology* 57 (2018), S104 – S92.
[67] C. Wah, Grant Van Horn, Steve Branson, Subhransu Maji, P. Perona, and Serge J. Belongie. 2014. Similarity Comparisons for Interactive Fine-Grained Categorization. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 859–866.
[68] B. Weijters, H. Baumgartner, and Maggie Geuens. 2016. The calibrated sigma method: An efficient remedy for between-group differences in response category use on Likert scales. *International Journal of Research in Marketing* 33 (2016), 944–960.
[69] Chris Welty, Praveen Paritosh, and Lora Aroyo. 2019. Metrology for AI: From Benchmarks to Instruments. *arXiv preprint arXiv:1911.01875* (2019).
[70] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. 1391–1399.
[71] S. Yan, H. Wang, T. Huang, Q. Yang, and X. Tang. 2007. Ranking with Uncertain Labels. *2007 IEEE International Conference on Multimedia and Expo* (2007), 96–99.
[72] S. Yan, H. Wang, X. Tang, Jianzhuang Liu, and T. Huang. 2008. Regression From Uncertain Labels and Its Applications to Soft Biometrics. *IEEE Transactions on Information Forensics and Security* 3 (2008), 698–708.
[73] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1005–1017. https://doi.org/10.1145/2818048.2819953
[74] Amy X Zhang, Jilin Chen, Wei Chai, Jinjun Xu, Lichan Hong, and Ed Chi. 2018. Evaluation and refinement of clustered search results with the crowd. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–28.
[75] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. 2017. Quantifying Facial Age by Posterior of Age Comparisons. *ArXiv* abs/1708.09687 (2017).

Received January 2021; revised April 2021; accepted May 2021