

"I Hope This Is Helpful": Understanding Crowdworkers' Challenges and Motivations for an Image Description Task

RACHEL N. SIMONS, Texas Woman's University

DANNA GURARI, The University of Texas at Austin

KENNETH R. FLEISCHMANN, The University of Texas at Austin

AI image captioning challenges encourage broad participation in designing algorithms that automatically create captions for a variety of images and users. To create large datasets necessary for these challenges, researchers typically employ a shared crowdsourcing task design for image captioning. This paper discusses findings from our thematic analysis of 1,064 comments left by Amazon Mechanical Turk workers using this task design to create captions for images taken by people who are blind. Workers discussed difficulties in understanding how to complete this task, provided suggestions of how to improve the task, gave explanations or clarifications about their work, and described why they found this particular task rewarding or interesting. Our analysis provides insights both into this particular genre of task as well as broader considerations for how to employ crowdsourcing to generate large datasets for developing AI algorithms.

CCS Concepts: • **Information systems** → **Crowdsourcing**; • **Computing methodologies** → **Computer vision**; **Computer vision tasks**; **Image representations**; *Machine learning*; • **Human-centered computing** → *Accessibility*.

Additional Key Words and Phrases: Crowdsourcing; Computer Vision; Artificial Intelligence; Image Captioning; Accessibility; Amazon Mechanical Turk

ACM Reference Format:

Rachel N. Simons, Danna Gurari, and Kenneth R. Fleischmann. 2020. "I Hope This Is Helpful": Understanding Crowdworkers' Challenges and Motivations for an Image Description Task. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 105 (October 2020), 26 pages. <https://doi.org/10.1145/3415176>

1 INTRODUCTION

An increasing number of services have emerged over the past decade to provide greater access to the wide range of visual information that surrounds us, often in the form of image descriptions or captions (we use these terms interchangeably). However, a serious challenge for these services is how to handle the enormous number of images that exist. Accordingly, while many captioning services are provided by remote humans [1, 3, 99, 111], the artificial intelligence (AI) community has recently begun developing algorithms that generate such image captions automatically as a faster, cheaper, and more scalable solution.

Successfully delivering automated captioning services is predicated on establishing large-scale datasets for training and evaluating the algorithms. Accordingly, over a dozen publicly-shared

Authors' addresses: Rachel N. Simons, rsimons@twu.edu, Texas Woman's University, 304 Administration Dr., Denton, Texas, 76204; Danna Gurari, danna.gurari@ischool.utexas.edu, The University of Texas at Austin, 1616 Guadalupe St, Suite 5.202, Austin, Texas, 78701; Kenneth R. Fleischmann, kfleisch@ischool.utexas.edu, The University of Texas at Austin, 1616 Guadalupe St, Suite 5.202, Austin, Texas, 78701.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2020/10-ART105 \$15.00

<https://doi.org/10.1145/3415176>

captioning datasets [6, 27, 28, 32, 36, 38, 47–49, 57, 58, 85, 91, 105, 106, 112] have been created over the past decade. The primary aim for the introduction of new datasets is to train algorithms to handle a greater diversity of both images and use cases.

Captioning datasets are often built with great involvement from crowdworkers recruited from Amazon Mechanical Turk (AMT) [32, 49], the world's largest crowdsourcing¹ platform. On this platform, individuals and organizations ("requesters") can post a variety of human intelligence tasks ("HITs" or "assignments") for anonymous workers ("Turkers") [23] to complete. The crowdworkers are hired to support dataset creation by creating the descriptions of images.

Our goal in this paper is to better understand the perspective of the Turkers who are generating the captions for these large-scale datasets. We expect that these Turkers could offer significant insights into both what is going well and what is going poorly with such captioning tasks—if only there were a way to let us know.

Accordingly, we collected 1,064 comments from Turkers who completed an image captioning HIT. Turkers were shown images taken by people who are blind and told that their image descriptions were intended to "help" people who are blind. Our thematic analysis of their comments is guided by the research question: *What are crowdworkers' reactions to creating image descriptions?*

Our findings reveal several limitations in the current widely-used approach for crowdsourcing image captions. Most limitations emerged due to our novel use case of assisting people who are blind to recognize content in self-taken images. Turkers experienced challenges both in understanding what people who are blind want described and how to handle unique characteristics of images taken by people who are blind (e.g., many are low quality [30]). We also identify the open-endedness of the captioning task as a significant challenge. We offer recommendations for how to better specify the image captioning task for Turkers, as well as for better supporting the specific use case of people who are blind in learning about their self-taken pictures.

In the following sections, we begin by reviewing the relevant literature. Next, we describe our dataset and the qualitative methods that we used to analyze it. We then report the results of this analysis and the broader implications of our findings for image captioning dataset creation for AI and using crowdsourcing tasks to support people who are blind. Finally, we share our conclusions.

2 RELATED WORK

We begin by reviewing related literature about services for delivering image captions to real users, as well as on approaches for developing large-scale datasets to support the development of image captioning algorithms. We then discuss an important gap in the existing literature on image captioning tasks: understanding and improving the experiences of crowdworkers.

2.1 Image Captioning Services for Real Users (i.e., People Who Are Blind)

Numerous image captioning services support people who are blind to learn about visual information. Some such services employ fully-automated solutions to describe digital images, such as those embedded in social media platforms (e.g., Facebook [35, 109] and Twitter [97]) and in productivity applications (e.g., Microsoft's PowerPoint [65]). Other services rely directly on humans to describe visual content [1–3, 16, 89, 99, 111]; for example, by asking captioners to describe digital images that people who are blind have taken [12]. These human captioners range from heavily-trained workers (e.g., Aira [2]) to untrained volunteers (e.g., Be My Eyes [4]).

¹While the term "crowdsourcing" is frequently used to encompass a number of different practices [33], crowdsourcing can broadly be understood as an "online, distributed problem-solving and production model" [14].

While it is clear that real users (i.e., people who are blind) desire image captioning services [94], only a few studies have focused on understanding the motivations of human captioners to support these services [17, 63]. Accordingly, our research complements prior research by revealing crowdworkers' perspectives of performing this important work.

2.2 The Critical Foundation of Image Captioning Algorithms: Large-Scale, Human-Annotated Datasets

A key inspiration for the development of image captioning algorithms is to improve the overall accessibility of images for people who are blind. Currently, the standard approach for developing such algorithms is to train them to describe images as a human would by showing them many human-captioned images. Such "training" datasets often are built with crowdsourced labor.

Many of the popular captioning datasets in the AI community were created using the same basic crowdsourcing task design. This task design, first developed in 2013 [49, 106], remains the standard approach [7, 28]. One concern about crowdsourced datasets built using this standard task design is that captions for the same image generated by different people can vary considerably [53, 98]. This variance makes it difficult to know what to teach algorithms and how to evaluate algorithm-generated captions.

While prior research into the issue of caption variance has focused on quantitatively characterizing its prevalence (e.g., [53, 98]), our qualitative analysis of crowdworkers' feedback about the captioning task provides a finer-grained understanding of *why* the variance may occur. In doing so, our findings are situated in the broader landscape of crowdsourcing research that seeks to better understand "inter-worker" variability (e.g., [8, 42, 90]). This understanding is a valuable precursor to supporting collaborative work and decision-making that can mitigate—or even support—such variability [24, 26, 54, 101].

More generally, to our knowledge, no prior work has examined feedback from crowdworkers about this standard captioning task design. Yet, gathering and analyzing worker comments to iteratively improve task design has long been encouraged within the broader crowdsourcing community (e.g., [71, 73]). Our work fills this gap by summarizing crowdworkers' direct comments about completing the image captioning task and by suggesting improvements to the task design.

2.3 Defining a "Good" Caption

A necessary part of training and evaluating captioning algorithms is establishing a definition for what makes an image caption "good" or successful. As mentioned above, some definitions assume that a "good" caption is one where multiple captions produced by independent people for the same image are similar (i.e., low variance) [53, 98]. Other definitions offer relatively vague guidance, including: accurately describe the image [32, 49, 59, 60, 64, 76], be grammatically correct [32, 60, 76, 104], lack incorrect information [76], be creatively constructed [64], and seem human-like [76].

While some definitions focus specifically on creating "good" captions for people who are blind [9, 11, 22, 79, 84, 94, 95, 100, 108], only a few studies directly integrate preferences reported by people who are blind [94]. To our knowledge, our work is the first to identify crowdworkers' questions and concerns about how to create good image captions—especially for people who are blind.

2.4 Understanding and Integrating the Experiences of Crowdworkers

While our work complements prior efforts to crowdsource image captions, our work is more broadly situated in the research about understanding the complexities with crowdsourcing (e.g., [56, 62, 73]).

2.4.1 Understanding Crowdworkers' Motivations. Previous research has examined crowdworkers' reasons for performing various crowdsourcing tasks [80, 92, 110]. Workers are frequently motivated by extrinsic factors such as payment, job-market signaling, competence development, and fostering social affiliation [83]. Intrinsic motivations can also be a significant factor in workers' decisions to select specific tasks [83] and to submit more work [23]. For example, workers performing citizen science and scientific crowdsourcing tasks are often highly motivated by intrinsic factors [23, 34, 68, 69]. Workers frequently balance different motivations simultaneously, such as the external motivation to make money and the internal motivation to develop one's creative skills [15].

Our work is most related to prior work that examined workers' motivations around providing captions (specifically intended to support people who are blind) for images embedded in digital learning resources [63]. This work focused on comparing only two motivating factors for captioners: monetary pay and altruism. In contrast, we present a range of motivations that crowdworkers described for creating image captions for people who are blind. Additionally, we focus on understanding workers' motivations for a novel task that consists of 1) describing images taken by people who are blind, in order to 2) explicitly contribute to the development of AI algorithms that can automatically caption images.

2.4.2 Social and Ethical Issues in Crowdsourcing. Several researchers have emphasized the need to more deeply consider the "ethics and values of crowdwork" [51], especially as it becomes a significant source for invisible and underpaid labor [39, 45, 51]. This area of research has produced a rich qualitative description of Turkers' experiences [20, 44, 61, 70, 102], including through ethnographic [40, 71, 72, 88] approaches. This work highlights issues such as power imbalances between the higher social and socioeconomic status that requesters have compared to Turkers [41] and Turkers' need to develop task-specific skills while "on-the-job" so that their work will be compensated [29, 55, 74].

Our work builds on such prior research and calls for designers to more closely attend not only to the design of *objects* such as crowdsourcing tasks and captioning algorithms, but also to the wider experiences of the *people* who use them [52]. In doing so, our findings complement recent research focused on encouraging requesters to collaborate directly with crowdworkers in order to improve task design [18, 67]—in our case, specifically for improving the task design for building large-scale datasets to support algorithm development.

3 DATASET

We now describe our dataset of crowdworkers' feedback about completing an image captioning task. We begin by describing the AMT HIT we deployed to simultaneously crowdsource image captions and Turkers' feedback about the task. Then, we describe our deployment of this HIT and post-processing of the Turkers' comments. Finally, we characterize our resulting dataset of Turkers' comments that we use for our qualitative analysis.

3.1 Collection of Captions with Comments

Observing that people who are blind are the primary audience for image captioning services, we sought to develop a novel large-scale captioning dataset that represents a real use case for these users. As discussed in Section 2.1, people who are blind currently share self-taken images in order to learn about their visual surroundings. Accordingly, our HIT engaged Turkers in captioning images taken by people who are blind in order to better design algorithms for this use case.

3.1.1 Basic Task Design for Crowdsourcing Captions. To generate this large-scale dataset of captions paired with Turker comments, we adapted the most popular image caption collection framework in

the AI community (discussed earlier in Section 2.2). The task includes the following basic elements: presentation of one image to caption at a time; a set of instructions on how to "describe the image," including a short list of things to do (e.g., include at least 8 words in the description) and things NOT to do (e.g., speculate about what people in the image might be saying or thinking); and a text entry box for workers to input their image descriptions.

3.1.2 Our Additions to the Basic Image Captioning Task Design. We designed our captioning task to support the use case of providing people who are blind with captions for images they take. A screenshot of our final task design is shown in Figure 1.

Our task design was developed iteratively through four pilot studies, in consultation with accessibility experts at Microsoft. We introduced several modifications to the standard task design. First, workers were notified at the beginning of the HIT that "[their] work will help to build smart systems that can automatically describe our visual world to people who are blind." We intended this text to provide insight into the motivation for our task. Second, we augmented the instructions with the following guidance: "describe all parts of the image that may be important to a person who is blind. E.g., imagine how you would describe this image on the phone to a friend." Third, workers could use a pre-formulated description to indicate when image quality issues made it "impossible to recognize the visual content (e.g., image is totally black or white)."

3.1.3 Caption and Comment Collection on AMT. To accelerate collection for our large-scale caption dataset, we included five images per HIT, with "next" and "previous" buttons allowing Turkers to move between images in the task. To improve the overall quality of the captions, we implemented a number of quality control mechanisms (summarized in Appendix section A).

For a random subset of assignments, we asked the following yes/no question: "Do you have close friends or family who are blind who you describe visual things to? (Your response will not affect the approval or rejection of your work.)" **Importantly for this work, for every assignment, we included an open-ended question asking Turkers if they "have any suggestions, feedback, or general comments for us." These comments were collected to better understand workers' experiences and, as discussed below, ultimately served as the foundation for our analysis in this paper.**

In total, we crowdsourced five captions per image for 39,213 images [43]. Our collection of images came from users of the VizWiz mobile phone application [12]; this application was designed to empower people who are blind to submit self-taken images (paired optionally with questions) in order to receive image captions or answers from remote humans.²

3.2 Crowdwork Overview

3.2.1 Characterization of Turker Involvement. In total, 1,623 unique individuals completed at least one HIT. The cumulative time for this task was 3,736.04 person-hours (the equivalent of over 93 40-hour work weeks) and the captioning duration was 101.52 hours (i.e., 37 hours of work completed every hour). The median time for an individual worker completing the HIT (based on all assignments from all batches used to create the final dataset) was 235.322 seconds (3 minutes and 55.322 seconds).

We used this median to calculate the average pay for a worker on our HIT. The projected hourly rate is: $(0.55 \text{ USD/HIT} * 3600 \text{ sec/hr}) / (235.322 \text{ sec/HIT}) = 8.41 \text{ USD/hr}$. As such, our average rate of pay is well above the standard average rate for similar crowdsourcing tasks [46, 87]. Of course,

²Photographers provided permission for their images to be used in dataset creation. No images with potentially private visual information were included.

[Hide / Show Details](#)

Motivation: Your work will help to build smart systems that can automatically describe our visual world to people who are blind.

We ask you to: carefully review images taken by people who are blind, and then (1) describe the image as per the instructions, (2) say if there is any text in the image, and (3) select what (if anything) is wrong with the photographic quality of the image.

PLEASE NOTE: It is possible that some images could be meaningless, inappropriate, or offensive. This is because we cannot control what pictures are taken. Kindly use your best judgement for this task.

IMPORTANT: Please do not refresh the webpage once you have started working, as you will lose all your progress, and have to start at the beginning.

[Hide](#)

You can see this information anytime by clicking "Hide / Show Details" button above.

Instructions:

Step 0: Please review the image. You may adjust your view using the toolbar:

- Use the + and - buttons (or your mousewheel) to **zoom in** or **zoom out**.
- Click and drag on the zoomed image to **pan around**.
- If needed, **rotate** the image using the last two buttons on the toolbar.


Step 1: Please describe the image as per the given prompts.

Step 2: Please indicate whether the image contains any form of text. Partial texts should also be reported.

Step 3: Please select what are the photographic **quality issue(s)** in the image.

- You may select more than one issue, or the "no issues" option.

Image 1
Image 2
Image 3
Image 4
Image 5
Finish



+
-
🔄
66%
🖼️

Step 1: Please describe the image in one sentence.

- Describe all parts of the image that may be **important to a person who is blind**.
E.g., imagine how you would describe this image on the phone to a friend.
- DO NOT** speculate about what people in the image might be saying or thinking.
- DO NOT** describe things that may have happened in the future or past.
- DO NOT** use more than one sentence.
- If text is in the image, and is important, then you can summarize what it says.
DO NOT use all the specific phrases that you see in the image as your description of the image.
- DO NOT** describe the image quality issues. This is covered in Step 3.
If the image quality issues make it **impossible to recognize the visual content** (e.g., image is totally black or white), then use the following description (you can copy-paste):

Quality issues are too severe to recognize visual content.
[Copy to description](#)

- Your description should contain at least **8 words**.

Type here. Do not start the description with:

- "There is/are ..."
- "This is / These are ..."
- "The/This image/picture ..."
- "It is/ It's ..."

Step 2: Is there any text in the image?

☐ YES: The image contains text.
☐ NO: The image has no text.

Step 3: What are the quality issues in the image?

☐ **BLUR:** Is the image blurry?

☐ **BRIGHT:** Is the image too bright (e.g., light is directly behind the object)?

☐ **DARK:** Is the image too dark (e.g., poor lighting in a room)?

☐ **OBSTRUCTION:** Is the scene obscured by the photographer's finger over the lens, or another unintended object?

☐ **FRAMING:** Are parts of necessary items missing from the image?

☐ **ROTATION:** Does the image need to be rotated for proper viewing?

☐ Any other issues: anything else not covered above

OR

☐ **NO ISSUES:** There are no quality issues in the image.

[Previous](#)
[Next](#)

Fig. 1. Screenshot of the user interface used to crowdsource the collection of image captions

some workers will have taken significantly longer to complete this HIT, making their rate of pay noticeably lower. (We discuss this more in the Results and Appendix sections B.)

3.2.2 Caption Consistency. As discussed in Section 2.2, achieving a high consistency between captions for the same image remains a significant concern in using crowdsourced datasets to train AI algorithms. Overall, we observed a similar level of consistency between captions in our dataset as in the most popular captioning dataset [65] (which was built with the same basic crowdsourcing task design). More details on caption consistency are provided in Appendix section D.

3.3 Crowdworkers' Comments

In total, 577 workers (roughly one third of all workers) submitted a total of 1,064 optional comments in response to our open-ended question asking workers if they "have any suggestions, feedback, or general comments for us." In the dataset prepared for our qualitative analysis, each comment was linked with an anonymous ID (one single ID per each individual worker) but no other worker information was included. Additionally, neither the exact images that each worker worked on nor their generated captions were linked to the comments. This dataset was then compiled into a .csv file with the comments grouped together by worker ID, in descending order by number of comments provided by each worker. (Figures showing the number of comments per worker and number of words per comment are discussed in Appendix section C.)

4 METHODS

We next discuss our qualitative approach to thematically analyzing workers' comments, including the contextualized set of questions that we developed to guide our analysis.

We used inductive thematic analysis [19] to qualitatively analyze all 1,064 comments in relation to our research question: *What are crowdworkers' reactions to creating image descriptions?* This method can be especially valuable when first approaching an under-researched topic or when approaching data without the use of formalized theory [75]. Additionally, thematic analysis may be particularly appropriate for approaching qualitative data when researchers have some prior sense of relevant concepts but would still prefer to allow for the discovery of additional, contextualized themes [13].

In accordance with the first step of this approach, therefore, we conducted a preliminary review of the comments to understand their level of detail and the topics that workers discussed. Only 6 comments appeared to be some kind of error or were otherwise uninterpretable and another 14 were apparently random and/or intentionally unrelated to the task. (The latter were all from the same worker and consisted of phrases such as "Goldeneye for N64," "Build the Wall," and "Hi mom.")

Drawing on the related work described above and our overarching research question, we then developed a set of contextualized questions for organizing our analysis of these comments:

- What challenges do workers express with understanding how to complete this task?
- What suggestions do workers have for improving this task?
- What kinds of explanations or clarifications do workers give about their work?
- Why do workers find this task rewarding or interesting?

In line with previous adoptions of this approach for analysis [37, 66, 81], two authors independently and inductively coded all comments at the comment level for both semantic and latent meaning. These authors then iteratively refined this list of codes together into themes connecting across the comments, then shared these themes with the other authors for finalization. We adhered to Braun and Clarke's [19] 15-point checklist of criteria for good thematic analysis to improve the validity of our results.

This process successfully generated several significant qualitative themes. We have chosen to focus only on those themes most closely related to workers' subjective experiences of the captioning task in this paper (and have set aside themes about challenges with flagging image quality issues and identifying the presence of text in images for future work).

5 RESULTS

We now discuss the results of our thematic analysis of worker comments. We identified key themes in relation to each of our contextualized questions, as well as some additional themes. In presenting these qualitative results, we provide a description of the predominant or important themes and include worker comments that illustrate these themes. Following previous work [74, 81], we additionally include counts (the number of comments that were coded under that theme) for several themes and sub-themes. We include these numbers "for the sake of illustrating internal coherency," rather than "as representative or used for inference" [81].

5.1 What Challenges Do Workers Express With Understanding How to Complete the Task?

Many comments discussed specific challenges that workers experienced with the task. These challenges are broken down in sub-themes below.

5.1.1 Using the Task Interface. Workers discussed several issues with using the task interface itself. In particular, several comments discussed issues related to manipulating and viewing the images such as the ease of correctly zooming in and out on the images (5 comments) or of rotating or panning the images (2 comments).

Workers also identified "missing," "blank," "empty," or "nothing" images (mostly images that failed to load) as an understandably significant challenge to completing the task (18 comments). For example, one worker commented that "there should be some way to report images that do [not] load."

5.1.2 Navigating the Quality Control Restrictions. As discussed above, our task design included a number of quality control mechanisms that enforce task requirements that are similar to previous crowdsourced image captioning tasks. Many comments addressed challenges with understanding and navigating these restrictions. Workers were particularly frustrated with the inability to use periods for purposes such as serving as decimal places (as periods were disallowed due to the single-sentence restriction). In total, 23 comments discussed this limitation in terms of being unable to describe prices, decimals in numbers, abbreviations, web sites, and proper titles.

Similarly, several workers discussed the challenge of adhering to the minimum word length, for example:

There may be images where there's important information that can be gleaned, but it may be hard to get to 8 words without adding filler words that you say should not be included.

One comment summed up the tension between the imposed limits of word count and sentence structure, saying, "The guidelines to have at least 8 words, but not more than a sentence, are silly." In total, 14 comments discussed similar challenges in meeting this length requirement.

5.1.3 What Kind of Detail to Include. Workers also discussed challenges in determining which specific types of visual information they should be including in their descriptions. Workers expressed uncertainty about how to describe images including for colors (12 comments), light direction, nutrition information (2 comments), off-frame information (2 comments), and other sight-based

adjectives such as visual texture ("chrome"). Several comments additionally referenced the perspective either of objects within the image ("Should the orientation of everything in the picture in relation to one another be described?") or of the image more generally ("Should perspective be indicated, such as looking down or from a high vantage point?").

In order to help them determine which such information was relevant, many workers discussed a desire for additional context about the task itself. For instance, 2 comments discussed wanting to know whether the users of these images were totally blind or had low vision. Workers were particularly interested in knowing more about the purpose of the images or in receiving other context that might inform them about what level of detail would be appropriate (16 comments).

5.1.4 Handling Unfamiliar Content. Workers additionally expressed some challenges about how to indicate that they are uncertain about their interpretation of an image's content. (This is distinct from being unable to describe an image due to image quality issues.) For example, one worker directly asked how to handle reporting this within the description, saying:

Is it possible to put something like, "A small blue suitcase is shown with WHAT LOOKS LIKE A BED to the left"? How do we mention something that we can't tell what it is, but we know the main part of the picture?

Another worker asked,

Can we put something like, "What looks like a brown and white dog....", if we're not sure it's actually a dog or an animal?

These examples highlight the subjective nature of both interpretation and uncertainty in creating image captions.

5.1.5 Handling Potentially Objectionable Content. Finally, 5 comments discussed specific issues that workers had with what they felt was objectionable content. One worker apparently objected strongly to the presence of insects in the images, saying:

I would appreciate if you did not show bugs, spider, roaches, or other insects in the photos. That is gross!

Another worker commented:

Image 5 in this series shows a pretty nasty foot parasite; not sure if that's something you want to include or not. It didn't bother me, but it might bother someone else.

Other comments discussed the presence of (primarily illustrated) nudity and additionally suggested that workers should be warned about "adult content" in advance.

5.2 What Suggestions Do Workers Have For Improving This Task?

Workers not only discussed challenges and uncertainties related to completing the task, but also made suggestions about how the task might be improved. Below we have broken down such suggestions into sub-themes.

5.2.1 Include Positive Examples. Many workers expressed a desire for extra guidance on how to successfully complete the task. The most common such suggestion involved the researchers providing more (and specific) examples before starting the task. For example, one worker suggested:

It would have been good if you had given a photo along with a sample description, so [that] we have a better idea of [if] what we are writing complies with what you are looking for. Thanks!

Of these comments, 21 directly referenced a need for more examples specifically of the desired sentence structure. For example, one worker suggested:

It would be helpful to have a sample descriptive sentence to base our descriptions off of. It is difficult to start a sentence without the phrases "It is, this is a photo of, This is," etc.

In total, 104 comments referenced a desire for examples (or samples) of successfully completed tasks. Additionally, several comments emphasized that it would be particularly helpful to specifically include some positive examples for workers to model their own descriptions on, rather than only providing a list of "too many rules" of what *not* to do:

It told me what not to start with ("a picture of," for instance) but it was really vague on how to start the description.

Several comments additionally indicated that including both positive and negative examples in the task description could alleviate concerns about the risks involved in accepting the HIT again.

5.2.2 Provide Feedback to Workers on Their Work. Several comments expressed a desire for workers to receive more feedback after completing the HIT, especially so that their work can be most helpful or acceptable (19 comments). For example:

I would appreciate feedback on how I did on this HIT so that I can give good quality data in the future, including things I could have improved on. Thanks!

Several such comments indicated that workers were interested in specifically repeating this task but were uncertain that they would accept the HIT again without receiving feedback.

5.2.3 Other Suggestions to Improve the Task. Workers also provided a variety of additional suggestions for improving the task that addresses many of the themes discussed above. Several comments expressed a desire for more options within the interface for easily indicating different issues that were not directly related to "image quality", including indicating when an image is unrecognizable to the worker and flagging "inappropriate" content.

Similarly, one comment suggested that the task "should note what to do [if] no image loads at all" and other suggested adding "an extra area to differentiate images that are potentially broken from the images that are severely low quality."

Finally, several workers had suggestions about how to address issues with the sentence structure limitations imposed by the quality control mechanisms. For example, one comment suggested that the task be re-designed to "change the 'one sentence' rule to a character limit" while another comment suggested that the task should "use 3 sentences: Describe the foreground. Describe the midground. Describe the background. Light direction. Makes more sense than just trying to cram it all into 1 long run-on sentence."

5.3 What Kinds of Explanations or Clarifications Do Workers Give About Their Work?

Workers frequently included additional contextual or explanatory information about their thought process while completing the HIT. These explanations and clarifications are primarily focused either on providing explanations related to their description of specific images or on clarifying their previous knowledge of or experience with people who are blind.

5.3.1 Image-specific Explanation. Several workers attempted to clarify what some object in an image might be (6 comments). For example, one worker commented:

The last image- I hate to use that copy [text] saying I can't identify it, but I can't tell what it is. I'm not sure if we're supposed to guess; because if I had to guess I would say maybe it's a TV screen with a video game on it, but I'm not sure if that would just be making something up.

While this example brings together many of the themes discussed above, it additionally demonstrates the intention of such explanatory comments to elucidate specific decisions made by the worker in the face of uncertainty.

Other such comments clarified why or how something had been labeled as text or *why* workers felt they did not know what a specific item was exactly. These voluntary, explanatory comments highlight workers' extra efforts to understand the task and to have their responses be understood (and accepted).

5.3.2 Clarification Related to Helping People Who Are Blind. As we mentioned previously, some workers were asked if they "have close friends or family who are blind who [they] describe visual things to." Several workers appear to have left additional clarifying information about their response to this question (14 comments). One worker specifically discussed the subjectivity in answering this question, saying:

I had a professor in college who I would help very rarely so I was putting "yes" to the above question, but I'm just realizing that was a long time ago and I only did that a few times. So I'm going to start answering "no" now.

An additional 4 comments clarify workers' personal opinion on what might be useful to people who are blind, for example:

When I was in college, I met a woman who was legally blind. She could identify skin color. It was quite remarkable! Thus, the description should help the blind person to imagine the color of the object. This can be really encouraging for those who lost their sight after many years of enjoying images.

In this case, rather than asking for additional guidance on what kinds of details to include in the description (as in the examples in the sub-section above), the worker is discussing what kinds of information *should* be included. Interestingly, this worker's previous experience helping a person who is blind does not seem to involve a close friend or family member.

Several workers similarly discussed their personal experience helping a person who is blind, but who was not directly related to them (7 comments). Some workers discussed their experience with progressively losing vision or having low vision themselves (5 comments). For example, one worker specified, "I myself am blind in one eye (from an inherited eye disease), with poor vision in the other," while another commented, "I am slowly losing my sight and thank you for the work that you are doing in these surveys." Several comments also discussed workers' experiences with having a family member who is blind (10 comments). This kind of personal connection or interest in people who are blind additionally impacted workers' motivations and interest in this task (as we discuss below).

5.4 Why Do Workers Find this Task Rewarding or Interesting?

Many comments seemed to comment positively about the task; overall, 81 comments expressed some variation of finding the task to be generally "good," "nice," or "wonderful." Many workers, however, were more specific about what they did or did not like about the task. We discuss themes related to workers' perceptions of the task and their motivations for selecting this particular task below.

5.4.1 What Workers Liked About This Task. In total, 59 comments described workers "liking" the task, or finding the task "satisfying," or "enjoyable." Another 14 specifically discussed the task as "fun," while 56 comments described the task as "interesting." Several other comments addressed the images as a source of enjoyment, with 8 comments describing the images as "good" or "nice," and 10 comments specifying a "like" for the specific content of the images.

In addition to liking the task itself, some workers emphasized that the task was a "good idea," or that they "like the idea" (14 comments). One worker described the task as a "good language challenge" and another felt that the task appealed to their skillset "as a writer." Similarly, 2 comments discussed the value of having "learned something" while completing the task and another 2 comments describe the task as a "nice change of pace."

Finally, 5 comments discussed liking that the task was "simple," while one worker commented that the pay for the task "is good."

5.4.2 What Workers Didn't Like About This Task. Workers also discussed several aspects of the task that they disliked or otherwise found unpleasant. Several workers felt that the task was "not as easy" as they had initially expected (4 comments), or that the task was actually "kind of hard" (3 comments). Similarly, workers commented that they felt that the length of time that they had to spend on the task was too great. One worker explained this issue in detail, saying:

There are too many images in one HIT, and it is very hard to get your head wrapped around what is expected and how to explain what is in the images to someone without sight. I believe that you will get a lot of people working very hard if you split up the images so [that] people are not taking so much time to complete more difficult ones, which will be dragging down the [overall] time. I spent nearly 25 minutes on this one HIT. I hope that you can adjust something to make it a bit more worthwhile. Thanks!

This example also highlights how taking excess time can reflect the heavy cognitive load that some images bring or the struggle that some individuals may have with tasks that are more open-ended.

Related to concerns about the length of the task is Turkers' sense of fair pay for their work. In total, 10 comments discussed workers' perceptions of being underpaid, especially in consideration of other elements of the task. One worker summed up this tradeoff, saying:

This HIT is grossly underpaid for the work time. Also [it is] difficult to describe things with so many rules [in place], [which] also puts a fear of rejection when you have so many rules. [T]hus, I'm only going to do 1 for now to see if I [did] it right.

In total, 8 comments directly addressed the fact that rejection is risky for workers. One comment particularly describes workers' tension of wanting to participate in this particular task, but being afraid of rejection:

I want to continue working on these tasks, but I'm not quite sure if I'm doing it right or meeting expectations. It makes it risky for Turk workers to accept the job because if it is rejected, our approval rating goes down. I was willing to take the risk because I love what this project is doing and want to be a part of it, but I do see the undesirable aspects that may come from these hits based on a lack of understanding in the instructions.

This example additionally highlights how many workers described being specifically interested in (and deliberately selecting) this task.

5.4.3 Why Workers Chose This Specific Task. Beyond describing aspects of the task that they liked and disliked, workers also commented on aspects of this particular task that appealed to them or made them want to continue. For example, some workers commented that they were "proud" of participating in this task (2 comments), or that they found the task to be an "honor" or "humbling" (2 comments). Other workers felt that the task is "important" (6 comments) or that the work is "useful" (5 comments). More specifically, some workers commented that they would like to "help people", "feel good" to help people (5 comments), or that they were "glad to help" (8 comments). One worker additionally commented that they wanted to "promote inclusion."

Beyond these somewhat more general motivations, many workers directly related their motivation for this task to the project's focus on people who are blind. In total, 16 comments discussed workers' appreciation for being able to help people who are blind. One worker commented:

This task is quite interesting! I work with blind people daily and I like knowing this might help with new technologies for them.

Similarly, another worker commented:

I'd love to know a description about more of what this project is for. Is it related to "Be My Eyes"? I love this; wish my grandmother had been around for this.

In addition to expressing a personal connection to the task, this worker indicated interest in knowing more about the specifics of the research project.

Finally, a deeply personal connection can be a powerful motivator for completing these tasks with a deep level of care and specificity. As one worker described in detail:

This project means a lot to me—raised by an artist grandmother, and caregiving for a Papa with no frontal vision and fading peripheral vision. [...] How do I know that the air duct in picture 2 is for heating and not air conditioning? Because I have spent most of my life in construction, as did Papa. Describing something to someone over the phone that you know and that has the same basic knowledge you do is much different than describing something to a stranger, blind or not. Another friend's father was blinded while working as a carpenter. He and I have much the same basic knowledge (I went into Electrical), but what about the random person walking into a gallery? Just how much detail do you want? (PS—love the racecar in picture 5 "Home Depot"—I'm there nearly every day.)

As is the case with this example, such lengthy and detailed comments were most common when workers also discussed having some personal connection or experience helping people who are blind. Moreover, this comment highlights the additional satisfaction that crowdworkers may feel when they are able to apply specific domain knowledge or work experience to the captioning task.

Workers additionally discussed being personally interested in doing more such tasks (5 comments) and wanting to know more about this project (3 comments). Additionally, 3 comments directly invited the researchers to contact them to follow up or to share more tasks.

5.5 Other Themes

A total of 25 comments contained miscellaneous references to specific image content, such as "I would like to purchase one box of such biscuit." Several workers expressed that they had "tried very hard" or "tried their best" (15 comments), and a couple of comments asked if workers could do "more than one" HIT. Additionally, as many of the above comments indicate, workers were generally quite polite. A total of 184 comments thanked us in some way, while 21 comments wished us luck with the project and another 11 simply said hello or wished us a good day/night.

Finally, 55 comments expressed a theme of concern or the expression, "I hope this is what you are looking for" or "I hope this helps." Other variations include: "hope I'm doing this right," "hope this works," "hope this is what you need," "I hope I helped," "hope this is helpful," "I'm hoping you get results that you want to find out of us doing these for you," and "I hope my descriptions help someone!" We combined these similar comments into a single theme because it was often impossible to determine who the intended "someone" was who was hopefully being helped—just as it was difficult to determine who the descriptions were hopefully "working" for—requesters versus people who are blind.

6 DISCUSSION

We now discuss some implications of our findings for captioning images for people who are blind, creating large-scale datasets for the AI community, and crowdsourcing to support accessibility.

6.1 Creating Image Captions for Real Users (e.g., People Who Are Blind)

Our findings in Section 5.1 highlight a gap in how some Turkers understand the interests of users who are blind. Although many Turkers wanted to provide caption information that would be specifically useful for people who are blind, they did not always know what information to include. Some of the clarifications requested by Turkers about how to create good captions parallel cutting-edge research (summarized in Section 2.1). In particular, our findings support recent studies examining the level and type of detail that blind users want in image captions for images found online [78, 94].

Our findings also expose challenges for human captioners that have not been discussed in prior work. In particular, Turkers were interested in knowing how to describe aspects such as color, nutrition information on food labels, relational context between objects observed in an image, the vantage point in which the picture was taken, and how much inferred knowledge to share about what cannot be seen in a poorly-framed image.

Accordingly, we believe a valuable direction for future work is to directly interview users of captioning services for the blind about what makes a caption most useful to them for *their own images*. With such knowledge, AI researchers might refine caption collection tasks to better address these interests and, ultimately, create algorithms that better serve these real users. Such efforts would reinforce the aims of user-centered (e.g., [5]) and participatory design (e.g., [93]), which have long advocated for greater inclusion of diverse users within design processes.

6.2 Addressing Issues with the Image Captioning Dataset Creation Process

As discussed in Section 2.2, we adapted our captioning task design directly from a task design introduced in 2013 [49] that has since been employed by many research groups [7, 28, 49, 106]. Our findings (specifically, Sections 5.1 and 5.2) reveal crowdworkers' frustrations with this task design and guidance for how to improve it.

We believe that some of our findings about the limitations of the captioning task design emerged because, unlike prior work, we employed a novel set of pictures taken by real-world users of captioning services. Previous work typically used a constrained collection of pre-filtered images curated from the internet, around a relatively limited set of topics [7, 28, 49, 49, 106]. One possible difference of our dataset is that many images contain content that requires incorporating decimal points to describe (e.g., thermometers and money). Turkers' comments revealed a need to distinguish between decimal points and periods; initially, our task design enforced a quality control mechanism that limited the caption to one sentence by permitting one period at most (but the task did not distinguish between periods and decimal points). Another distinction of our HIT is that Turkers likely sought more regular and advanced use of image manipulation tools (e.g., image rotation, panning, zoom) in order to make sense of the image content *because* many images were lower quality (since users who are blind cannot verify the quality of their images).

Our findings offer strong evidence that the basic captioning task design, which has seemingly become the status quo for caption collection, will need to continuously evolve in order to be inclusive of the diverse types of visual data that real users could want captioned. As a first step to support this evolution, we will publicly share the code to our final crowdsourcing system design along with the crowdsourced captions and Turkers' anonymized comments:

<https://vizwiz.org/tasks-and-datasets/image-captioning/>. We expect that new limitations of the task design will continue to emerge as researchers expand the focus of image captioning tasks to include a more diverse collection of images, including: memes, data visualizations, technical figures, anime, pictures from robots, and pictures taken using wearable devices.

Our own experience with adapting a well-established task design reinforces the importance of recent work within the crowdsourcing research community that advocates for collaboration across stakeholders in order to accomplish effective, iterative task design for crowdsourcing tasks [18]. In our case, four pilot studies with random samples from our image collection were not sufficiently comprehensive to capture the long tail of limitations that we would discover. Rather, new issues were identified through our continuous monitoring of worker feedback. Our findings underscore the importance for the AI community to update the status quo to include collecting and sharing worker feedback as an integral part of publicly-shared training datasets.

6.3 Instructing Crowdworkers for Dataset Creation Tasks for the AI Community

As discussed in the Dataset section, our task introduced instructions to the standard captioning task design indicating that crowdworkers would "help to build smart systems that can automatically describe our visual world to people who are blind." Our findings reveal that some workers connected meaningfully with this motivation, commenting that they *"like knowing this might help with new technologies for them."* However, such responses do not fully indicate the extent to which workers understood either the nature of dataset creation to improve image captioning algorithms or how such "smart systems" will be used by people who are blind—as exemplified by comments such as *"I'd love to know a description about more of what this project is for. Is it related to 'Be My Eyes'?"*.

In retrospect, some of the ambiguity found in such comments is likely related to a lack of clarity in who the intended target audience was. Telling workers that their work would "help to build smart systems that can automatically describe our visual world to people who are blind" does not give them much ability to understand either the algorithm development process or the needs of users who are blind. While the photo-takers and the caption-receivers for this research are theoretically the same group (users who are blind), there are actually multiple other user groups who utilize workers' captions: the immediate research team and the larger research community who will be using the captions. Workers likely were somewhat confused about what audiences and uses to tailor their captions to. This confusion is perhaps reflected in the frequency of comments that express "I hope this is helpful" or "I hope it helps someone."

Our findings underscore the need for a focused examination of the trade-offs of how much information to include in task instructions for crowdworkers creating datasets that will support algorithm development, including in areas beyond image captioning tasks. Potential areas of information include (1) providing more details about requesters' acceptance/rejection criteria; (2) explaining that workers' contributions will be used to build automated systems that might replace their annotation efforts in the future; (3) clarifying how such future automated systems differ from existing services (e.g., Be My Eyes); and (4) detailing for whom such automated systems are being built (e.g., providing personas of target users). While much previous work on designing crowdsourcing tasks has focused on (1) (e.g., [74]), our findings suggest that the other areas of information would benefit from greater consideration when designing tasks intended to leverage crowdworkers to support AI development (e.g., via A/B testing).

Such strands of consideration should examine both ethical and practical trade-offs that are relevant to the broader AI and crowdsourcing research communities. For instance, is it ethical to employ workers in contributing to automated systems that may ultimately replace their efforts without informing them of this potential outcome? What are the appropriate levels of detail and explanation for conveying the motivation of a dataset creation task in order to mitigate confusion

from some crowdworkers? These considerations are particularly applicable to the many researchers who are creating large-scale, human-labeled datasets to support algorithm development.

6.4 Considering the Trade-offs of Open-Ended Captioning Tasks

Our findings also highlight that part of Turkers' confusion about our HIT stems from its open-ended nature. While an open-ended approach to this captioning task design has been used since 2013 [49], the nature of this approach may cause workers to struggle unnecessarily with deciding what types of information to convey and what level of detail to provide. Accordingly, another way to reduce workers' confusion (in addition to clarifying the task instructions) could be to reduce the open-endedness of such tasks.

One solution may be to provide workers with a more specific list of visual information to discuss. Such close-ended approaches might potentially even go so far as to provide a fixed checklist instead of a free-text box. Alternatively, templates could be created that, for example, identify what types of content to insert when constructing a sentence [77, 107]. Such approaches "prime" workers towards certain work outcomes by giving them examples.

As an alternative to both open- and closed-ended approaches, task designers might consider developing novel "middle-ended" approaches to generating captions. For example, the task could begin by presenting workers with a small set of vignettes taken from user studies done with people who are blind. Such vignettes might describe different scenarios and uses for captions (including the appropriate level of caption detail) based on the experiences of people who are blind. Of course, this approach opens up a different challenge of how to reduce the wide diversity of visual interests of a user population into a few meaningful vignettes.

On a higher level, it is important to note that what makes a "good" caption for users who are blind may not be ultimately deemed synonymous with what is deemed a "good" caption for AI researchers (as discussed in section 2.3). Our findings support previous work indicating that workers' subjective, different interpretations may be desirable [8, 10, 25, 42, 78, 101, 103]).

6.5 Engaging Crowdworkers in Accessibility-Related Tasks

An aspect of our findings that we did not anticipate centers on the demographics of workers on AMT. While extensive literature discusses various social and economic demographic characteristics of Turkers [31, 46], our work is the first study (to our knowledge) to discuss the familiarity of the AMT crowd with the interests/needs of people who are blind. We were surprised to learn that many Turkers expressed intimate familiarity with this population (via marriage, profession, family, personal experience, etc.). Moreover, through this work, we initiated a continued correspondence with numerous workers who were eager to contribute to follow-up studies supporting people who are blind. Accordingly, crowdsourcing on AMT has proven to be a useful platform for recruiting allies of people who are blind to engage in research. This may complement previous work, which found that volunteer workers with a range of personal experience with assistive technologies contribute to collaborative efforts to design and improve such technologies [21, 50, 82, 86].

7 CONCLUSION

In the pursuit of automated AI solutions to scale up impact, we should not forget about the crowdsourced labor currently working "behind the scenes" to generate large datasets. While human crowdworkers are well suited to contribute, they understandably would like significant context for navigating the complexity of subjectivity in performing the tasks. Our findings are relevant both to improving the quality of the data generated by such tasks, as well as to improving the experience for workers. Better supporting crowdworkers in performing such tasks is not only an ethically and socially valuable goal in itself, but helping them to do better work can also create a virtuous cycle

[96] whereby workers will be encouraged (and will encourage others) to continue successfully performing similar tasks in the future.

Looking beyond the specific area of image captioning, the approach of integrating worker feedback as a necessary step in dataset creation can potentially be generalized to a variety of fields. Complementary ways for future work to solicit worker feedback include employing detailed ethnographic and observational studies of workers in order to generate real-time data about their experiences, as well as developing large-scale surveys to gain additional understanding of workers' experiences. Ultimately, refining the design of such tasks to better support workers' needs and goals can allow them to "do their best," especially when workers "hope [their work] is helpful."

ACKNOWLEDGMENTS

We gratefully acknowledge funding from Microsoft and thank the anonymous crowdworkers for their significant contributions to this effort. We also thank the anonymous reviewers, Meredith Ringel Morris, Ed Cutrell, and Besmira Nushi for their valuable feedback and discussions about this work.

REFERENCES

- [1] 2018. *BeSpecular*. <https://www.bespecular.com>
- [2] 2018. *Home - Aira*. <https://aira.io/>
- [3] 2018. TapTapSee - Blind and visually impaired assistive technology - Powered by the CloudSight.Ai Image Recognition API. <https://taptapseeapp.com/>
- [4] 2020. Be My Eyes. <https://www.bemyeyes.com/>
- [5] Chadia Abras, Diane Maloney-Krichmar, and Jenny Preece. 2004. User-centered design. In *Encyclopedia of Human-Computer Interaction*, W. Bainbridge (Ed.). Vol. 37. Thousand Oaks: Sage Publications, 445–456.
- [6] Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2018. Nocaps: Novel Object Captioning at Scale. *arXiv preprint arXiv:1812.08658* (2018).
- [7] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2018. Nocaps: Novel object captioning at scale. *arXiv:1812.08658 [cs]* (Dec. 2018). <http://arxiv.org/abs/1812.08658> arXiv: 1812.08658.
- [8] 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 (WWW '19)*. Association for Computing Machinery, San Francisco, USA, 1100–1105. <https://doi.org/10.1145/3308560.3317083>
- [9] Cynthia L. Bennett, Jane E. Martez E. Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, Montreal QC, Canada, 1–12. <https://doi.org/10.1145/3173574.3173650>
- [10] Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why Does a Visual Question Have Different Answers?. In *Proceedings of the IEEE International Conference on Computer Vision*. 4271–4280.
- [11] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 333–342. <https://doi.org/10.1145/1866029.1866080> event-place: New York, New York, USA.
- [12] Jeffrey P. Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz: LocateIt-Enabling Blind People to Locate Objects in Their Environment. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference On*. IEEE, 65–72.
- [13] Anne E. Bowser, Oliver L. Haimson, Edward F. Melcer, and Elizabeth F. Churchill. 2015. On vintage values: The experience of secondhand fashion reacquisition. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 897–906. <https://doi.org/10.1145/2702123.2702394> event-place: Seoul, Republic of Korea.
- [14] Daren C. Brabham. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence* 14, 1 (Feb. 2008), 75–90. <https://doi.org/10.1177/1354856507084420>
- [15] Daren C. Brabham. 2012. Motivations for participation in a crowdsourcing application to improve public engagement in transit planning. *Journal of Applied Communication Research* 40, 3 (Aug. 2012), 307–328. <https://doi.org/10.1080/>

00909882.2012.693940

- [16] Erin Brady. 2015. Getting fast, free, and anonymous answers to questions asked by people with visual impairments. *SIGACCESS Access. Comput.* 112 (July 2015), 16–25. <https://doi.org/10.1145/2809915.2809918>
- [17] Erin Brady, Meredith Ringel Morris, and Jeffrey P. Bigham. 2015. Gauging Receptiveness to Social Microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1055–1064. <https://doi.org/10.1145/2702123.2702329> event-place: Seoul, Republic of Korea.
- [18] Jonathan Bragg, Mausam, and Daniel S. Weld. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. ACM, New York, NY, USA, 165–176. <https://doi.org/10.1145/3242587.3242598> event-place: Berlin, Germany.
- [19] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.
- [20] Alice M. Brawley and Cynthia L. S. Pury. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54 (Jan. 2016), 531–546. <https://doi.org/10.1016/j.chb.2015.08.031>
- [21] Erin Buehler, Stacy Branham, Abdullah Ali, Jeremy J. Chang, Megan Kelly Hofmann, Amy Hurst, and Shaun K. Kane. 2015. Sharing is Caring: Assistive Technology Designs on Thingiverse. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, Seoul, Republic of Korea, 525–534. <https://doi.org/10.1145/2702123.2702525>
- [22] Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. 2012. Crowdsourcing Subjective Fashion Advice Using VizWiz: Challenges and Opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 135–142.
- [23] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (June 2013), 123–133. <https://doi.org/10.1016/j.jebo.2013.03.003>
- [24] 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 (CHI '17)*. ACM, New York, NY, USA, 2334–2346. <https://doi.org/10.1145/3025453.3026044> event-place: Denver, Colorado, USA.
- [25] 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 (CHI '17)*. Association for Computing Machinery, Denver, Colorado, USA, 2334–2346. <https://doi.org/10.1145/3025453.3026044>
- [26] 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 (CHI '16)*. Association for Computing Machinery, San Jose, California, USA, 3180–3191. <https://doi.org/10.1145/2858036.2858411>
- [27] Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. Déjà image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 504–514.
- [28] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325* (2015).
- [29] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. 2018. Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 37:1–37:17. <https://doi.org/10.1145/3274306>
- [30] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. 2020. Assessing Image Quality Issues for Real-World Problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3646–3656.
- [31] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of Mechanical Turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 135–143. <https://doi.org/10.1145/3159652.3159661> event-place: Marina Del Rey, CA, USA.
- [32] Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1292–1302.
- [33] Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science* 38, 2 (April 2012), 189–200. <https://doi.org/10.1177/0165551512437638>
- [34] Alexandra Eveleigh, Charlene Jennett, Ann Blandford, Philip Brohan, and Anna L. Cox. 2014. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2985–2994.
- [35] Facebook. 2015. Facebook: Milestones. <https://www.facebook.com/facebook?sk=info>
- [36] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*. Springer, 15–29.
- [37] Casey Fiesler and Blake Hallinan. 2018. "We are the product": Public reactions to online data sharing and privacy controversies in the media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI*

- '18). ACM, New York, NY, USA, 53:1–53:13. <https://doi.org/10.1145/3173574.3173627> event-place: Montreal QC, Canada.
- [38] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.
- [39] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt. Google-Books-ID: 8AmXDwAAQBAJ.
- [40] Mary L. Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. The crowd is a collaborative network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 134–147. <https://doi.org/10.1145/2818048.2819942> event-place: San Francisco, California, USA.
- [41] Neha Gupta, David Martin, Benjamin V. Hanrahan, and Jacki O'Neill. 2014. Turk-life in India. In *Proceedings of the 18th International Conference on Supporting Group Work*. ACM, 1–11.
- [42] 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*. ACM, 3511–3522.
- [43] Danna Gurari, Yanan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning Images Taken by People Who Are Blind. *arXiv:2002.08565 [cs]* (Feb. 2020). <http://arxiv.org/abs/2002.08565> arXiv: 2002.08565.
- [44] Benjamin V. Hanrahan, David Martin, Jutta Willamowski, and John M. Carroll. 2018. Investigating the Amazon Mechanical Turk market through tool design. *Comput. Supported Coop. Work* 27, 3-6 (Dec. 2018), 1255–1274. <https://doi.org/10.1007/s10606-018-9312-6>
- [45] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 449:1–449:14. <https://doi.org/10.1145/3173574.3174023> event-place: Montreal QC, Canada.
- [46] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Benjamin V. Hanrahan, Jeffrey P. Bigham, and Chris Callison-Burch. 2019. Worker demographics and earnings on Amazon Mechanical Turk: An exploratory analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, LBW1217:1–LBW1217:6. <https://doi.org/10.1145/3290607.3312970> event-place: Glasgow, Scotland Uk.
- [47] David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 237–244.
- [48] William Havard, Laurent Besacier, and Olivier Rossec. 2017. Speech-Coco: 600k visually grounded spoken captions aligned to Mscoco data set. *arXiv preprint arXiv:1707.08435* (2017).
- [49] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.
- [50] Jonathan Hook, Sanne Verbaan, Peter Wright, and Patrick Olivier. 2013. Exploring the Design of technologies and services that support do-it-yourself assistive technology practice. *Proceedings of DE 2013* (2013).
- [51] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 611–620. <https://doi.org/10.1145/2470654.2470742> event-place: Paris, France.
- [52] Lilly C. Irani and M. Six Silberman. 2016. Stories We Tell About Labor: Turkopticon and the Trouble with "Design". In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4573–4586. <https://doi.org/10.1145/2858036.2858592> event-place: San Jose, California, USA.
- [53] Mainak Jas and Devi Parikh. 2015. Image Specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2727–2736.
- [54] 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 (CSCW '16)*. Association for Computing Machinery, San Francisco, California, USA, 1637–1648. <https://doi.org/10.1145/2818048.2820016>
- [55] Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey P. Bigham. 2018. Striving to Earn More: A Survey of Work Strategies and Tool Use Among Crowd Workers. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*. <https://www.aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17920>
- [56] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. <https://doi.org/10.1145/1357054.1357127>
- [57] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? Text-to-image coreference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3558–3565.
- [58] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, and David A. Shamma. 2017. Visual genome: Connecting language and vision using crowdsourced dense

- image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.
- [59] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Babytalk: Understanding and Generating Simple Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2013), 2891–2903.
 - [60] Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg, and Yejin Choi. 2014. TreeTalk: Composition and Compression of Trees for Image Descriptions. *Transactions of the Association for Computational Linguistics* 2 (Dec. 2014), 351–362. https://doi.org/10.1162/tacl_a_00188 Publisher: MIT Press.
 - [61] Laura Lascau, Sandy J. J. Gould, Anna L. Cox, Elizaveta Karmannaya, and Duncan P. Brumby. 2019. Monotasking or multitasking: Designing for crowdworkers' preferences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 419:1–419:14. <https://doi.org/10.1145/3290605.3300649> event-place: Glasgow, Scotland Uk.
 - [62] Edith Law, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams. 2017. Crowdsourcing as a tool for research: Implications of uncertainty. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1544–1561. <https://doi.org/10.1145/2998181.2998197> event-place: Portland, Oregon, USA.
 - [63] A. M. Layas and Helen Petrie. 2016. Exploring intrinsic and extrinsic motivations to participate in a crowdsourcing project to support blind and partially sighted students. *Universal Design 2016: Learning from the past, designing for the future (Proceedings of the 3rd International Conference on Universal Design, UD2016)*. (Aug. 2016). <http://eprints.whiterose.ac.uk/118514/>
 - [64] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL '11)*. Association for Computational Linguistics, Portland, Oregon, 220–228.
 - [65] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft Coco: Common Objects in Context. In *European Conference on Computer Vision*. Springer, 740–755.
 - [66] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. "Could you define that in bot terms?": Requesting, creating and using bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3488–3500. <https://doi.org/10.1145/3025453.3025830> event-place: Denver, Colorado, USA.
 - [67] V. K. Chaithanya Manam and Alexander J. Quinn. 2018. WingIt: Efficient Refinement of Unclear Task Instructions. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*. <https://www.aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17931>
 - [68] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E. Schwamb, Chris J. Lintott, and Arfon M. Smith. 2013. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *First AAAI conference on human computation and crowdsourcing*.
 - [69] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why stop now? Predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.
 - [70] Catherine C. Marshall and Frank M. Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13)*. ACM, New York, NY, USA, 234–243. <https://doi.org/10.1145/2464464.2464485> event-place: Paris, France.
 - [71] David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14)*. Association for Computing Machinery, Baltimore, Maryland, USA, 224–235. <https://doi.org/10.1145/2531602.2531663>
 - [72] David Martin, Jacki O'Neill, Neha Gupta, and Benjamin V. Hanrahan. 2016. Turkling in a global labour market. *Comput. Supported Coop. Work* 25, 1 (Feb. 2016), 39–77. <https://doi.org/10.1007/s10606-015-9241-6>
 - [73] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (March 2012), 1–23. <https://doi.org/10.3758/s13428-011-0124-6>
 - [74] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2271–2282. <https://doi.org/10.1145/2858036.2858539> event-place: San Jose, California, USA.
 - [75] Lydia Michie, Madeline Balaam, John McCarthy, Timur Osadchiy, and Kellie Morrissey. 2018. From her story, to our story: Digital storytelling as public engagement around abortion rights advocacy in Ireland. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 357:1–357:15. <https://doi.org/10.1145/3173574.3173931> event-place: Montreal QC, Canada.
 - [76] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé. 2012. Midge: generating image descriptions from computer vision detections. In

- Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*. Association for Computational Linguistics, Avignon, France, 747–756.
- [77] Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hasty, and Steven Landau. 2015. Guiding novice web workers in making image descriptions using templates. *ACM Transactions on Accessible Computing (TACCESS)* 7, 4 (2015), 12.
 - [78] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 59.
 - [79] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "With most of it being pictures now, I rarely use it": Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, San Jose, California, USA, 5506–5516. <https://doi.org/10.1145/2858036.2858116>
 - [80] Babak Naderi. 2018. *Motivation of workers on microtask crowdsourcing platforms*. Springer, Cham, Switzerland. OCLC: 1020790439.
 - [81] Midas Nouwens and Clemens Nylandsted Klokmose. 2018. The application and its consequences for non-standard knowledge work. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 399:1–399:12. <https://doi.org/10.1145/3173574.3173973> event-place: Montreal QC, Canada.
 - [82] Jeremiah Parry-Hill, Patrick C. Shih, Jennifer Mankoff, and Daniel Ashbrook. 2017. Understanding Volunteer AT Fabricators: Opportunities and Challenges in DIY-AT for Others in e-NABLE. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, Denver, Colorado, USA, 6184–6194. <https://doi.org/10.1145/3025453.3026045>
 - [83] L. G. Pee, E. Koh, and M. Goh. 2018. Trait motivations of crowdsourcing and task choice: A distal-proximal perspective. *International Journal of Information Management* 40 (June 2018), 28–41. <https://doi.org/10.1016/j.ijinfomgt.2018.01.008>
 - [84] Helen Petrie, Chandra Harrison, and Sundeev Dev. 2005. Describing Images on the Web: A Survey of Current Practice and Prospects for the Future. *Proceedings of Human Computer Interaction International (HCII)* 71 (2005).
 - [85] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 139–147.
 - [86] Ludovico Orlando Russo, Giuseppe Airò Farulla, and Carlo Boccazzi Varotto. 2018. Hackability: A Methodology to Encourage the Development of DIY Assistive Devices. In *Computers Helping People with Special Needs (Lecture Notes in Computer Science)*, Klaus Miesenberger and Georgios Kouroupetroglou (Eds.). Springer International Publishing, Cham, 156–163. https://doi.org/10.1007/978-3-319-94274-2_22
 - [87] Susumu Saito, Chun-Wei Chiang, Saiph Savage, Teppei Nakano, Tetsunori Kobayashi, and Jeffrey P. Bigham. 2019. TurkScanner: Predicting the hourly wage of microtasks. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 3187–3193. <https://doi.org/10.1145/3308558.3313716> event-place: San Francisco, CA, USA.
 - [88] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We are Dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1621–1630. <https://doi.org/10.1145/2702123.2702508> event-place: Seoul, Republic of Korea.
 - [89] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2017. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. *Proceedings of HCOMP 2017* (2017).
 - [90] Mike Schaeckermann, 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 (Nov. 2018), 154:1–154:19. <https://doi.org/10.1145/3274423>
 - [91] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2018. Engaging image captioning via personality. *arXiv preprint arXiv:1810.10665* (2018).
 - [92] Luiz Fernando Silva Pinto and Carlos Denner dos Santos Júnior. 2018. Motivations of crowdsourcing contributors. *RAI: Revista de Administração e Inovação; São Paulo* 15, 1 (2018), 58–72. <http://search.proquest.com/docview/2063479696/abstract/648431A1613B4846PQ/1>
 - [93] Jesper Simonsen and Toni Robertson (Eds.). 2013. *Routledge international handbook of participatory design*. Routledge, New York. OCLC: 754734489.
 - [94] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. Honolulu, HI, USA, 13. <https://doi.org/10.1145/3313831.3376404>
 - [95] Abigale J. Stangl, Esha Kothari, Suyog D. Jain, Tom Yeh, Kristen Grauman, and Danna Gurari. 2018. BrowseWithMe: An Online Clothes Shopping Assistant for People with Visual Impairments. In *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*.

- [96] Rebekah Steele and Marjorie Derven. 2015. Diversity & Inclusion and innovation: A virtuous cycle. *Industrial and Commercial Training* 47, 1 (Jan. 2015), 1–7. <https://doi.org/10.1108/ICT-09-2014-0063>
- [97] Twitter. 2015. About Twitter, Inc. <https://about.twitter.com/company>
- [98] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.
- [99] Luis Von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu, and Manuel Blum. 2006. Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 79–82.
- [100] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How Blind People Interact with Visual Content on Social Networking Services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, San Francisco, California, USA, 1584–1595. <https://doi.org/10.1145/2818048.2820013>
- [101] Meihong Wang, Yuling Sun, Jing Yang, and Liang He. 2018. Enabling the Disagreement among Crowds: A Collaborative Crowdsourcing Framework. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. 790–795. <https://doi.org/10.1109/CSCWD.2018.8465368>
- [102] Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. 2017. "Our privacy needs to be protected at all costs": Crowd workers' privacy experiences on Amazon Mechanical Turk. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 113:1–113:22. <https://doi.org/10.1145/3134748>
- [103] Chun-Ju Yang, Kristen Grauman, and Danna Gurari. 2018. Visual question answer diversity. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.
- [104] Yezhou Yang, Ching Lik Teo, Hal Daumé, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Edinburgh, United Kingdom, 444–454.
- [105] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. Stair captions: Constructing a large-scale Japanese image caption dataset. *arXiv preprint arXiv:1705.00823* (2017).
- [106] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.
- [107] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual Madlibs: Fill in the Blank Description Generation and Question Answering. In *Proceedings of the Ieee International Conference on Computer Vision*. 2461–2469.
- [108] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 121.
- [109] Yuxiang Zhao and Qinghua Zhu. 2014. Effects of extrinsic and intrinsic motivation on participation in crowdsourcing contest: A perspective of self-determination theory. *Online Information Review; Bradford* 38, 7 (2014), 896–917. <https://doi.org/10.1108/OIR-08-2014-0188>
- [110] Haichao Zheng, Dahui Li, and Wenhua Hou. 2011. Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce* 15, 4 (July 2011), 57–88. <https://doi.org/10.2753/JEC1086-4415150402>
- [111] Yu Zhong, Walter S. Lasecki, Erin Brady, and Jeffrey P. Bigham. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2353–2362.
- [112] C. Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. 2013. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*. 1681–1688.

A CAPTION QUALITY CONTROL

Prior to the qualitative analysis conducted on the worker comments, several quality control mechanisms were employed to enhance the quality of the image captions. These quality control mechanisms are discussed in more detail in a related publication about the image caption dataset [43].

During the task, workers were prevented from submitting captions that did not meet the following criteria: (1) at least 8 words in the description, (2) not more than one occurrence of a period followed by a space (in order to restrict workers to providing a single, complete sentence), and (3) description does not begin with "There is," "There are," "This is," "These are," "The image," "The picture," "This image," "This picture," "It is," or "It's".

As an additional quality control mechanism, we manually reviewed a subset of the captions generated by specific workers. The qualifications used to determine which Turkers' captions to review include:

- Workers who were a statistical outlier in time-to-submit by 1.95 times the standard deviation of all results (i.e., taking either very little or very much time to complete the task)
- Workers who had CAPS LOCK enabled for more than half of the caption text
- Workers who used the pre-formulated text ("Quality issues are too severe to recognize visual content") for more than half of the images that they captioned
- Workers who were the only one to either use or to not use the pre-formulated text for a given image
- Workers who used words like "quality", "blur," and "blurry" (not including the pre-formulated text), and so may not have been focusing on the content in the image
- A random sample from all results

After such review, we rejected a total of 992 assignments that were submitted by 18 different Turkers. (A rejection also means that the Turker did not get paid for that assignment.) Each rejected assignment also received an explanatory message.

Of these rejected assignments, 548 were submitted by a total of 8 Turkers whom we considered to be potentially problematic and blocked from participating in subsequent tasks. These Turkers were given a message such as, "We're sorry to REJECT your work, as you constantly SPAMMED our HITs by entering garbage captions. We had no other choice but to BLOCK you from accepting our HITs in the future."

Due to the anonymization of the comment dataset and its separation from the worker captions, comments from these 8 Turkers may be included in our analysis; we decided not to distinguish whether or not any such comments were of lesser "quality" than other Turkers' comments. Accordingly, our final dataset of worker comments might still contain "spam" or "bot" submissions.

B CROWDWORKER EFFORTS

Figure 2 shows the number of images that workers took the same approximate amount of time to caption, with a noticeable number of images taking more than 220 seconds. Although most images apparently required less than 65 seconds to caption, a significant number of images required 2 or even 3 times more time (and presumably, more cognitive effort).

Figure 3 shows the distribution of the number of crowdworkers who completed a similar number of total HITs. The vast majority of workers completed fewer than 20 HITs, while the 31 workers completing over 200 HITs are apparently outliers (and may represent bot activity).

Together, these two figures can be used to gain an overall impression of worker effort on the captioning HIT, as well as distributed compensation for the task.

C FREQUENCY OF CROWDWORKERS' COMMENTS

Figure 4 shows the overall distribution of the number of comments with a certain number of words. While almost half of all comments (521) consisted of fewer than 7 words (median=7 words), the remaining comments likely contain more substantive feedback than would be provided by a bot. As shown in Table 1, most repeat commenters only commented twice and the most prolific commenter provided 128 comments (representing a significant outlier and potentially a bot).

Table 1. Number of Comments Made By Workers

Number of Workers	Number of Comments
404	1
110	2
26	3
16	4
4	5
7	6
3	7
1	9
1	10
1	11
1	12
1	19
1	26
1	128

D CAPTION CONSISTENCY

The "specificity score" measure assesses the similarity of captions generated by different human workers for the same image [53]. Each image is given a score between 0 and 1, with 1 indicating perfect caption consistency.

Figure 5 shows the number of images in our dataset with a particular specificity score. The 1,000 images clustered at the score of 1 are images for which all workers used our pre-formulated text. For comparison, Figure 6 shows the number of images in the MSCOCO dataset with a particular specificity score (only including MSCOCO Val Set due to the larger size of the dataset).

Received January 2020; revised June 2020; accepted July 2020

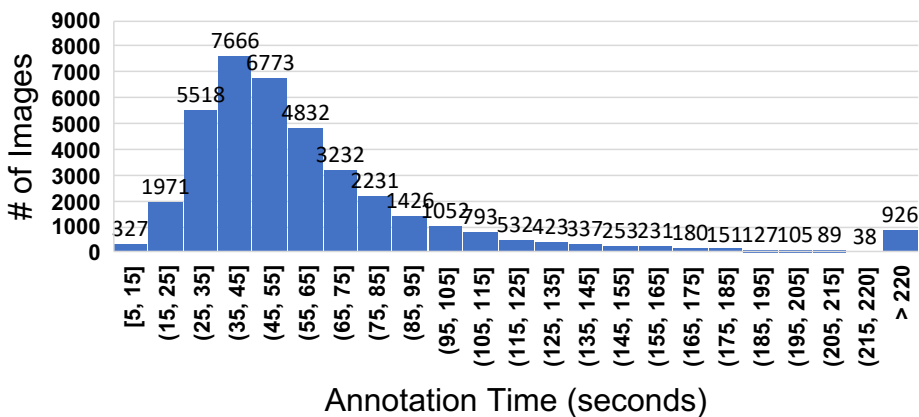


Fig. 2. Captioning Time Per Image

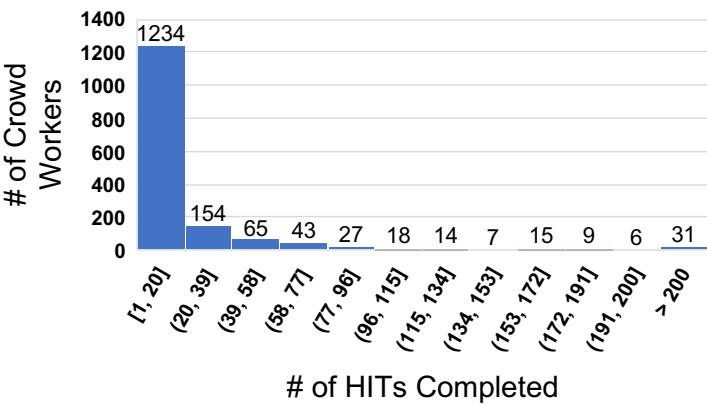


Fig. 3. Number of HITs Crowdworkers Completed

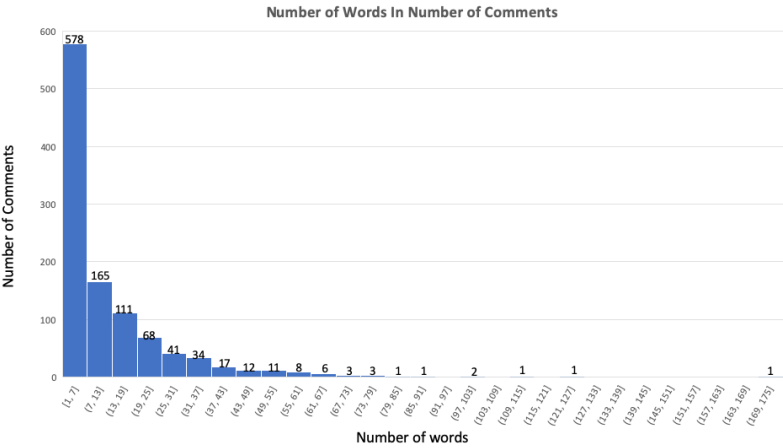


Fig. 4. Number of Words in Number of Comments

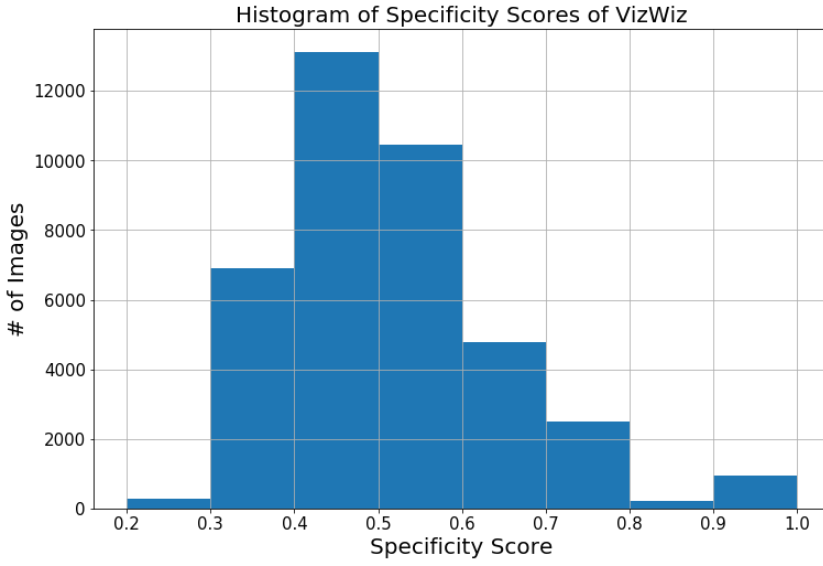


Fig. 5. Number of VizWiz Images with Range of Specificity Scores

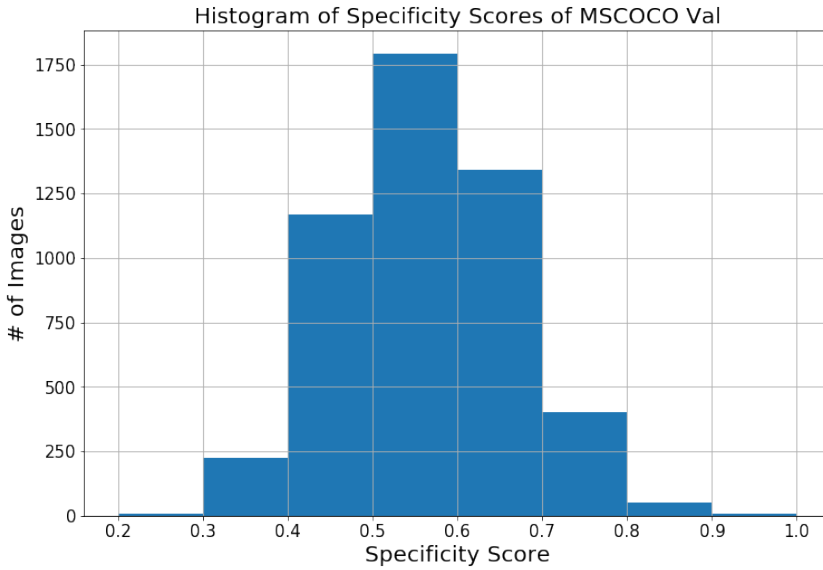


Fig. 6. Number of MSCOCO Images with Range of Specificity Scores