Deaf and Hard-of-Hearing Users' Prioritization of Genres of Online Video Content Requiring Accurate Captions

Larwan Berke, Matthew Seita, Matt Huenerfauth
Golisano College of Computing and Information Sciences, Rochester Institute of Technology
Rochester, NY USA
lwb2627@rit.edu, mss4296@rit.edu, matt.huenerfauth@rit.edu

ABSTRACT

Online video is an important information source, yet its pace of growth, including user-submitted content, is so rapid that automatic captioning technologies are needed to make content accessible for people who are Deaf or Hard-of-Hearing (DHH). To support future creation of a research dataset of online videos, we must prioritize which genres of online video content DHH users believe are of greatest importance to be accurately captioned. Our first contribution is to validate that the Best-Worst Scaling (BWS) methodology is able to accurately gather judgments on this topic by conducting an in-person study with 25 DHH users, using a card-sorting methodology to rank the importance for various YouTube genres of online video to be accurately captioned. Our second contribution is to identify video genres of highest captioning importance via an online survey with 151 DHH individuals, and those participants highly ranked: News and Politics, Education, and Technology and Science.

CCS CONCEPTS

• Human-centered computing~Empirical studies in accessibility.

KEYWORDS

Captioning, Video, Deaf and Hard-of-Hearing, Genres.

1 Introduction

We investigate the preferences of people who are Deaf or Hard of Hearing (DHH) as to which genres of online videos are most important to be accurately captioned. Online videos' importance as a communication medium has increased, yet the quantity of online video grows too rapidly for current captioning solutions to keep pace. There is currently **no legal mandate to caption all online video** in the U.S. [58] (especially user-generated content), and laws vary internationally. The DHH community has thus far been **dissatisfied with automatic captions** for online video based on automatic speech recognition (ASR) [59], and rigorous evaluation of such technology is needed.

With a prioritization from DHH users as to which genres of online video content are most important to have accurately captioned, companies with popular online video platforms could **focus investment** of human-authored or -corrected captioning, and investigate providing better accuracy for automatic captioning on high-priority genres. In addition, researchers investigating automatic methods to make videos accessible can use this genre-prioritization to assemble a testing/evaluation **video collection,** so that evaluation of their solutions are grounded in DHH users' interests. For this use, a prioritization of genres is more helpful than a simple binary classification as to whether each genre is important. ASR researchers can also assemble online videos (based on the genre-prioritization from DHH users), which can be accurately captioned by a human, to create training data for machine learning research, e.g. to investigate adaptation of ASR models for higher accuracy on specific video genres of interest to the DHH community, or for evaluation of current ASR systems on such videos.

In service of our main research focus on prioritizing various genres of online video, we also had to address a general methodological issue: *How can we gather rankings from DHH participants among a relatively large set of options, using a distributed survey methodology?* While someone in an in-person study may be encouraged to carefully rank a large number of items, there are known challenges in doing so using in an online survey [10, 55, 57]. While a technique called Best-Worst Scaling (BWS) [32] can support remote collection of high-cardinality ranking judgements, to our knowledge, no prior published research has ever used BWS with DHH users , who are known to have diverse levels of written language literacy [1, 35, 43]. For this reason, it was necessary to first conduct a rigorous statistical **validation** of the use of BWS through an in-person study with DHH users, before we determined that it was sound to deploy it in an online survey with a larger group of DHH participants.

The **contributions** of this study are as follows:

- **An empirical research contribution:** We conducted an online survey and in-person interviews to produce a **prioritization** of genres of online video that are most or least important to be accurately captioned, based on the collection of many judgements from the DHH community. Our research provides and discusses both **quantitative** (numerical rankings

W4A '20, April 20–21, 2020, Taipei, Taiwan
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7056-1/20/04...$15.00
https://doi.org/10.1145/3371300.3383337

of genres) and **qualitative** (subjective responses to open-ended interview questions) data.

- **Secondary, methodological contribution:** We validate the use of the BWS algorithm [32] to gather ranking preferences about online video genres from DHH users in an online survey. This is not only the first study to use BWS with DHH users, but we have conducted a **validation** that BWS results in equivalent findings as an in-person card-sorting task.

2 Background and Related Work

Text captions can make online video accessible for people who are Deaf or Hard-of-Hearing (DHH), who constitute a significant portion of the world's population: In the U.S., nearly one out of five people report some level of hearing loss [40]. Research has found that these individuals lack equal access to videos and television programming, which even when it is captioned, the captioning may contain errors. Captions, especially those that are automatically generated, can sometimes be missing words, delayed, difficult to read, or have other problems [2, 16].

Internet-based video is an increasingly popular platform for social-media engagement, entertainment, and education. Video social-media platforms have increased in usage over the past decade[1], and these platforms are increasingly important venues for political organizing, e.g. as in [53], and disseminating information about breaking news events [26]. Subscription video services, e.g. Netflix, are increasing in popularity[2], and multiple new online education platforms with video have been launched recently, e.g. Khan Academy or MIT OpenCourseware[3]. With the increasing importance of online video for participation in society, several researchers have begun to examine its accessibility, e.g. [9, 28, 29, 49].

As compared to television, standards for accessibility of online content are still emerging [47]. In addition to major network programming or original content produced by studios released via online platforms, there is a tremendous amount of user-generated content uploaded daily (e.g. over 400 hours of video is estimated to be uploaded to YouTube every minute[4]). Given the cost of professional captioning services, researchers have begun to examine technologies for automatically (or semi-automatically) providing captioning for such video, e.g. [18].

Given the diverse genres of video available online, it would be valuable for industry and researchers to know how to best prioritize their efforts at providing high-quality captioning for online content, or to use this information to produce research datasets (as discussed in the Introduction). While studies have examined the popularity of various genres of online video for a general audience, e.g. [6, 54], there is a gap in the literature: There has thus far been limited research on how to best prioritize which genres of online video DHH users are most interested in watching, and which of these genres they deem as highest-priority for receiving captions. While it would be ideal for all video content online to be fully captioned for these users, given the distributed nature of video generation (including user-uploaded videos) and resource constraints, it may be important to focus initial efforts on particular genres judged by the DHH community to be of high importance. While tracking current viewing patterns may seem like a proxy for identifying the importance of captioning various genres, it is not a perfect one: Lack of captions currently could prevent users from viewing videos they would like to watch, or some popular genres of video may have little aural information content necessitating captions.

As discussed below, prior work has focused primarily on television programming and DHH users' preference for the appearance of television captions. There has been less research on *online video captioning* for DHH users. Although some research has compared genres of video among a general audience, there is a gap in the literature: It is unknown what genres DHH individuals are most interested in watching and which genres they believe are more important to be captioned. One challenge in studying this issue is that prior work has found difficulty in asking survey participants to rank large sets of items. Below, we also describe an existing method, called Best-Worst Scaling (BWS) for efficiently collecting ranking data from participants, noting that no prior work has been done specifically using BWS with DHH participants, especially in an online survey context; so, its efficacy was unknown.

2.1 Prior Research on Caption Preferences

Prior research has investigated DHH users' captioning requirements [20, 21, 28, 29, 39]. Jensema *et al.* performed foundational work on DHH viewers' experiences watching television with captions [21]. Several researchers have studied how users balance their attention between the video content and captioning (as it can often be difficult to concentrate fully on video content at the same time as reading captions) [20, 29], and others have studied whether DHH users preferred captions or just reading a transcript [28]. Other work examined DHH viewers' perspectives on the importance of captions during advertisements [39].

Several groups have investigated the effect of subtle timing delays in captions on DHH users' experience [4, 27, 36]. Burnham *et al.* investigated whether out-of-sync captions impacted the viewing experience [4]. Bad timing was found to increase the DHH viewer's cognitive load [36], and too-fast captions caused frustration for DHH viewers [27].

Much of the literature has also focused on the best ways to present captions to DHH users [9, 11, 46, 52]. Researchers have examined whether captions should include "extra information," e.g. displaying applause, musical inserts, or other noise [11]. One team asked DHH users how they thought the captions should look on the screen, e.g. with or without a solid black background

[1] People watch over a billion hours of videos daily on YouTube. https://www.youtube.com/yt/about/press.

[2] Netflix doubled to 100 million members from 2014 to 2017. https://media.netflix.com/en/about-netflix.

[3] http://www.khanacademy.org and http://ocw.mit.edu.

[4] http://www.everysecond.io/youtube.

[46]. Another team made general recommendations about how captions should be located on the screen [9]. Finally, Automatic Speech Recognition (ASR) captioning is starting to show promise, and one group asked DHH users whether captions should include various types of punctuation [52].

Prior work has investigated whether DHH users would accept newly emerging Automatic Speech Recognition (ASR) captioning [3, 23, 45, 49]. Prietch *et al.* performed a large-scale literature review [45] and found that ASR had some success in providing captioning for DHH students in classrooms. One group asked DHH users which situations they would want to use ASR technology for online videos [49]. Researchers also developed methods to evaluate the usability of ASR captioning for DHH users [23]. Finally, some researchers have investigated using ASR for live captioning in one-on-one meetings [3].

Most research has examined captioning for television programming or some live contexts. The little prior work on online video has specifically focused on appearance or formatting issues. As discussed above, while there are advances in some automatic methods for captioning video, e.g. through ASR, such technology is imperfect. While providing high-quality captions for all online video content is an important future goal, in the near-term, given the volume of videos being uploaded to the internet every day, it is not feasible to accurately caption them all. Based on prior research, we now have knowledge on *how* to best display and format captions on online videos, and we know that ASR technology is not advanced enough to caption *every* video online with acceptable accuracy. As discussed above, since there is no legal mandate for all online videos to be captioned [58] and since the DHH community judge automatic captioning methods as insufficient [59], there is a need to understand *which* online videos DHH users care the most about having accurately captioned. In the short-term, priorities from DHH users can guide organizations in focusing human-powered efforts to provide high-accuracy captions. In the long-term, to drive research on improving automatic methods, video datasets (reflecting actual priorities of DHH users) can be constructed.

2.2 Prioritizing Video Genres for Captioning

While studio-produced video content on television channels or online streaming services may be produced or curated by the creators into various genres (sports, comedy, etc.), when considering online-video content more broadly, categorizing the videos into genres is non-trivial.

As a starting point for identifying a list of genres of video, we can consider how videos have been categorized in prior research. For instance, prior work on television captioning [19, 20] has considered genres such as *Films, News, Documentaries, Talk shows, Soap operas, Sports, etc.* Other groups of researchers used machine classification methods [38, 51] to automatically identify video genres such as *Newscasts, Cartoons, Football, Music, Weather, etc.* Some researchers interested in DHH viewers used eye-tracking [7] or perception questions [15] to explore genres such as *Action Movie, Documentary, Culture, etc.* More recently, hearing college students' usage of online video platforms [6] included these genres: *Reality show, Entertainment magazine, Education/how to, etc.* Finally, a large collection of YouTube videos was analyzed [54] in order to improve the indexing of search engines and included genres such as: *Autos & Vehicles, People & Blogs, Travel & Events, etc.*

However, there has been little prior work to understand what types of videos are popular among the DHH community. More research is needed to identify what DHH people like and what genres of videos they want to be captioned. As discussed above, it is important to note that statistics on current viewership of various online videos is not the best proxy - just because something is already popular to watch *for the general population*, there is a potential that this is different than what DHH users want to watch. Perhaps what DHH users want to watch is **not yet captioned** (so they cannot watch it without missing a lot of information), or perhaps DHH users are simply interested in **different things** than what hearing users are interested in. Furthermore, there is a difference between what people *want* to watch and what they think is *important to be captioned.* For instance, we could speculate that some genres (e.g. "videos of animals or pets") may be of interest, but perhaps a lack of spoken or aural content in such videos would mean that they are judged as **lower priority** by DHH users as to how important it is that captions be added. Looking at the prior literature, we do not see prior work in this space, and thus we have identified a gap in the literature that we attempt to fill in this study.

2.3 Challenges in High Cardinality Ranking

Given the diversity within the DHH community, we would prefer to conduct an online survey with many participants, yet since we wish to obtain a ranking among a large number of video genres, we foresee challenges. Prior HCI research has found that ranking a set of items is difficult due to issues such as self-deception, memory effects, and ordering effects [42, 50, 55, 56]. With such a large list of genres to rank (16), it would be challenging for participants to answer our question **reliably**. While card-sorting or drag-and-drop ranking is feasible with in-person studies, the online questionnaire literature recommends limiting forced-rank questions to a few items [57]. Further, if a question item for obtaining a complete ranking is burdensome or lengthy, there is risk that participants in an online survey may not be motivated to complete such a question item carefully.

Prior work has investigated methods for collecting ranking judgments from study participants more efficiently in studies, including a technique called "**Best-Worst Scaling**" (BWS) which was first proposed by Louviere & Woodworth in [33] and further described in a paper from Finn & Louviere [12]. The approach used by BWS nicely avoids the common pitfalls of ranking by asking participants to pick the "best" and "worst" items in an N-tuple (list of items, usually N=4) subset of the master list (M), and repeats the question many times (rounds) in order to include all of the items in M. BWS allows the participant to easily compare a small list of genres (the N-tuple) through several rounds of best/worst annotation instead of being cognitively overwhelmed by a single question with all of the items to rank, as could happen during traditional "integer ranking" methods (wherein the participant views the list of all items then gives a numerical rank to each item or sorts them into ascending/descending order, thereby implying the same numerical rank). This reduction in cognitive load would alleviate many of the challenges discussed in the previous paragraph in regard to obtaining rankings of a large number of items from DHH participants. Different BWS methods have been used successfully in several fields such as: Healthcare [13, 37], Social Sciences [8, 44], Agriculture [22, 48], and Natural Language Processing [24, 25]. This technique, however, has not been validated in the context of DHH users engaged in an online survey, which is a contribution of our paper.

In addition, prior research has measured diverse levels of written language literacy among members of the DHH community [1, 35, 43], and it is unknown how these users may interact with complex questions for ranking of items in high-cardinality sets. If there were a misunderstanding of a question prompt, then there could be risk that the responses gathered would not be a true measure of user's preferences. In theory, shorter questions, each with fewer answer choices, may be easier for someone with lower literacy to respond to; there may be less risk of information overload on any one question. Our study investigates whether these BWS questions work well with DHH individuals. In consideration of these literacy factors, our online-survey deployment of BWS with DHH users will include both English and ASL instructions, and to evaluate the reliability of BWS with these users, we compare BWS-based ranking results to those obtained from an in-person card-sorting activity conducted with both English and ASL instructions.

In summary, we are aware of **no prior published research that has ever used BWS among DHH individuals** (not during in-person studies nor during remote surveys). Given the unique literacy and language context of DHH users, it was unsafe to assume that BWS would work (or to make claims based on its output) without any prior validation.

3 Research Questions and Methods

Despite the importance of online video and the challenges in accurately captioning this fast-growing content, our literature review did not reveal prior work that had methodically studied the preferred genres of the DHH community when it concerns captioning. A recent trend in the field of computing accessibility is to consider the diversity within the DHH community by including as many participants as possible in studies gathering requirements or preferences, e.g. through online surveys [3]. Using the same rationale, it would benefit the research community to investigate this topic with a larger pool (over 100) of DHH participants and ask them for their opinion on the genres of captioned online videos. In the remainder of this paper, we examine the following questions:

RQ1: Does an online Best-Worst Scaling survey with DHH participants yield a ranking of a large number of video genres that is similar to a ranking obtained via an in-person card-sorting by the same participants?

RQ2: What thresholds in BWS scores could be used to label video genres as high, medium, or low importance?

RQ3: In an online BWS survey, which genres of online videos do DHH individuals believe are the most/least important to be accurately captioned?

RQ4: In open-ended comments, how do DHH users explain their prioritization of various video genres?

3.1 Overview: Summary of Our Studies

We first conducted an in-person interview study (see **In-Person Card-Sort and BWS**) with 25 DHH participants, who shared their opinions about the importance of 16 video genres in three different ways: answering BWS questions, performing a card-sorting of all genres, and grouping genres into levels (high, medium, and low importance). We had two goals from this study: (**RQ1**) By comparing the BWS results and card sorting, we determined if they resulted in similar ranking results (so we could confidently use BWS in a subsequent online survey). (**RQ2**) By comparing the three-level grouping to BWS, we determined how scores from BWS could be interpreted categorically, by identifying thresholds that partition our ranked list of genres into high, medium, and low importance levels.

Next, we conducted an online study using BWS with 151 DHH participants (see **Online Survey**) to prioritize video genres (**RQ3**). Finally, we returned to the data obtained in our interview study: We conducted a thematic analysis of responses to open-ended interview questions related to the importance of captioning various video genres (**RQ4**).

3.2 Identifying Genres to Rank

One approach for gathering judgments about important genres of online video to caption would be to ask users to suggest genres in open-ended questions. While this method may be useful to catching genres early in a project, it is difficult to quantitatively merge preferences among a large group of users. Different individuals may partition videos into different groupings, which may not be easily aligned in a one-to-one manner. For a large survey study, there is a need to provide users with a list of genres, to obtain ranking preferences. Our team used the list of 16 genres presented in this section for all studies described in this paper, but to check if we had missed any genres of importance to the DHH community we asked an open-ended question in our interview study (see **In-Person Card-Sort and BWS**). Participants listed genres they often watched (before seeing our list of genres) and their responses included a variety of TV-style genre types (*Action*, *Drama*, *Fantasy*, etc.) and more online-focused ones (*LGBT*, *RPG Games*, *Makeup/How-To*, etc.). After comparing our list of 16 genres with the 308 suggestions from interview respondents (73 unique), we did not find genres that were not already subsumed by our existing list.

Furthermore, there is an advantage to using a categorization scheme for online videos that aligns with that of major online-video companies; this may increase the utility of our findings for industry as they prioritize captioning efforts. Some prior research [5, 54] had made use of a pre-existing classification scheme of video genres published by Google YouTube[5]. As DHH participants often mentioned *YouTube* in prior captioning studies (e.g. P88's comment in [3]), it was logical for our team to use this set of labels as a starting point. However, we noticed that "Videoblogging" had been deprecated from this list of genres: *YouTube* renamed it as "People and Blogs." Considering the importance of "vlogs" (video blogs) to the DHH community [17], we decided to append that genre to *YouTube's* list. Fortuitously, for our Best-Worst Scaling procedure (described in the next section), it was elegant for our list to be a multiple of 4: Our final list of genres (16) that we use through this project is:

Animals and Pets, Autos and Vehicles, Comedy, Education, Entertainment, Film and Animation, Games, How-To and Style, News and Politics, Non-Profits and Activism, Music, People and Blogs, Sports, Technology and Science, Travel and Events, Video-Blogging.

3.3 Best-Worst Scaling (BWS) Instrument

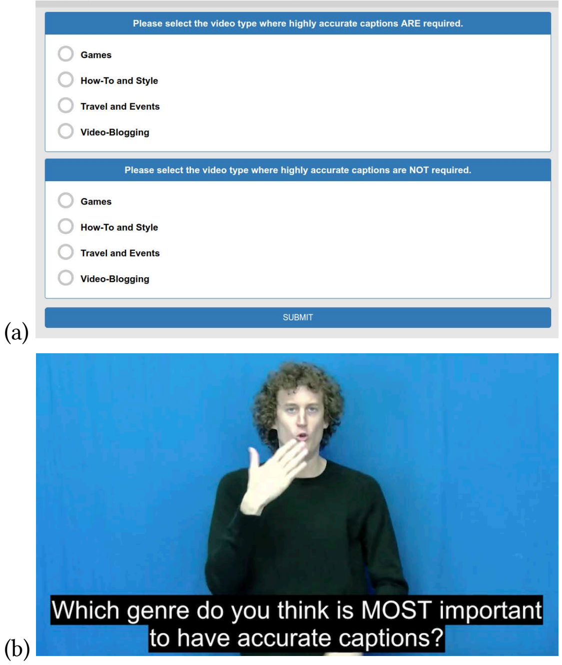
Since it can be overwhelming to ask someone to rank a set of 16 items, especially on an online survey question, we needed a method of presenting the genres to participants. Throughout this paper, we have used an online dynamic version of one of the BWS methods described on page 14 of [32].[6] Our participants would be presented N-tuples (subsets of our 16 genres shown as answer choices), and they would select the single **most** and single **least important** item from each N-tuple, in response to the question: **Which online video genre is the most/least important to be accurately captioned?** Across multiple rounds of questions, they would repeat this procedure with different N-tuples (subsets of the 16 items), see Figure 1.

Typical deployments[7] of BWS use an off-line program to generate the entire set of N-tuples *in advance* for the participant to rate. Whenever using BWS, an expansion factor (E, usually ~2x) is applied to the master list of items so different combinations of the items are selected for each N-tuple. The reason for this "expansion" is in order to satisfy three general rules for BWS, as on page 17 of [32]:

R1) In an N-tuple, there are no identical items.

R2) Each item is shown to the participant equally often.

R3) Each pair of items are shown equally often.



Please select the video type where highly accurate captions ARE required.

○ Games
○ How-To and Style
○ Travel and Events
○ Video-Blogging

Please select the video type where highly accurate captions are NOT required.

○ Games
○ How-To and Style
○ Travel and Events
○ Video-Blogging

SUBMIT

(b)
Which genre do you think is MOST important
to have accurate captions?

Figure 1. (a) BWS example with "Best" on top and "Worst" on bottom. (b) ASL video instructions with English subtitles.

In our case, setting E=2 for 16 genres would force our participants to answer 32 rounds of questions, but pilot testing revealed this would be too lengthy. As our research question (most/least important genre) is focused on the "ends" of the ranking, we made use of a variation of the BWS procedure: Rather than randomly choosing items for the N-tuples (which may require participants to spend time making fine-grained ranking decisions among mid-ranked genres), we instead used a dynamic online variant of BWS (as opposed to utilizing a static list generated off-line). In this way, we could present viewers with tuples focused on the most/least important genres based on the participant's prior answers. Before participants answered the BWS questions (example in Figure 1), we provided instructions in both ASL and English. Online supplementary files are included (available at http://latlab.ist.rit.edu/w4a2020) containing the ASL video used for these BWS instructions, to enable replication of our work by future researchers. An English transcript is provided below (the content was first authored in ASL and then translated to English):

> *Now, we want to learn what types/genres of videos you watch online. We will give you a list of genres, and you need to think carefully before answering. Which genre do you think is MOST important to have accurate captions? (I will be frustrated if the captions are bad for this genre) Pick one. Then, which genre do you think is LEAST important to have accurate captions? (I won't care if the captions are bad for this genre) Pick one. With your choices set for the MOST important and LEAST important genres, you can then answer the question. We will ask you this question several times with a different list of genres each time!*

[5] We downloaded the list on May 31, 2018 from https://support.google.com/youtube/answer/4594615, links to http://gdata.youtube.com/schemas/2007/categories.cat.

[6] Specifically, we used the BWS Object Case or "Case 1," rather than Case 2 (profile) or Case 3 (multi-profile) used in some other work.

[7] Kiritchenko and Mohammad performed NLP studies [24, 25]: http://saifmohammad.com/WebPages/BestWorst.html.

```
// First phase
itemList = shuffle( master item list ) N =
number of items in a tuple while ( !
empty itemList ) {
  db_store( p_ID, display( array_pop( itemList, N ) ) ) }
// Second phase
while ( db_get_num_ties( p_ID, best || worst ) != 0 ) {
  bestTied = db_get_ties( p_ID, best ), db_get( p_ID,
    lowest_freq_items )
  worstTied = db_get_ties( p_ID, worst ), db_get( p_ID,
    lowest_freq_items )
  db_store( p_ID, display( array_pop( bestTied, N/2 ),
        array_pop( worstTied, N/2 ) ) )  }
```

Figure 2. Pseudocode for dynamic BWS used in this paper.

Pseudocode for the dynamic BWS variant we used is provided in Figure 2, and it is briefly described here: Assume there are M items in the entire set being ranked and we are showing participants N-tuples at a time; the algorithm has two phases and iterates until it runs out of items to display to the participant. During the **first phase**, the M items are partitioned randomly into N-tuples, and the participant goes through all of the items to provide their initial opinion. Thus, there are M/N rounds in the first phase. At the end of this phase, some items will have received a best (most important) vote from a participant, some items will have received a worst (least important) vote, and the majority of items would have never received any vote. The **second phase** consists of several additional rounds of N-tuples that are shown to the participant. As we are interested in the most/least important genres, this phase acts as a focusing lens, forcing them to make a choice between pairs of items they previously had selected as best or worst. In addition to breaking remaining ties among the "best" and "worst" items, the algorithm also includes some items from the "middle" group from the end of the first phase, to satisfy BWS rules $\mathbf{R_2}$ and $\mathbf{R_3}$ and to balance the mix of items in each N-tuple (so it includes some items that had not received any best/worst votes, to reduce the risk of a participant giving an item a best vote in one round and a worst vote in another (thus cancelling each other out and delaying the algorithm's termination). At the conclusion of this algorithm, for all of the M items in the entire set (in our case, the 16 video genres), we have the following: Each item will have received some number of "best" votes (most important video genre in an N-tuple) and some number of "worst" votes (least important in N-tuple). From this, the BWS literature [32, 14] explains how to produce a score for each item on a scale from [-1,1], with higher values indicating "best" (most important video genre). Items can be sorted using these scores into a ranked list.

4 Organization of Methods and Results Below

The upcoming sections of this paper are organized according to our four research questions, with each section interleaving the methods, results, and a discussion of that question. We begin with a formal validation of BWS among DHH users via an in-person study (RQ1), followed by determining how to interpret BWS scores to identify the most/least important genres (RQ2). Next, we perform a large online survey using BWS to identify the most and least important genres (RQ3), followed by an analysis of open-ended responses from participants to understand why some genres were prioritized by these users (RQ4).

5 RQ1: BWS Validation with DHH Users

In this RQ, we investigate whether BWS can be used in studies with DHH participants. We pilot tested our online questionnaire: 5 Deaf participants (age 34-90, 3 female, 2 male) responded to the instrument online (remotely), and an in-person session included 6 Deaf, 1 deaf, and 1 Hard-of-Hearing participants (ages 21-27, 4 females). The pilot tests revealed practical issues that other researchers conducting computerized surveys with DHH users may want to know:

- Our participants used a variety of devices (laptops, smartphones, desktops - with a variety of OS); so, our survey UI had to be responsive across platforms. We also had to convert the informational videos into three versions (mp4, webm, and ogv) in order for our HTML5 player to function properly for everyone.
- We had included some informational videos (in ASL) which also displayed with English captions, but some participants told us it was insufficient because the captions were too fast or it was visually distracting. They preferred to view a transcript of the English instructions below the video; so, we provided: video in ASL, English captions on video, AND a transcript.
- We added an "Instructional Manipulation Check" in the survey UI to check for careless/inattentive participants as recommended in [10, 41].

After completing the pilot studies, we revised our online questionnaire instrument, which is used in studies below.

5.1 In-Person Card-Sort and BWS

To evaluate whether responses to BWS survey questions could be used to determine ranking preferences from these users, we recruited 25 DHH participants (18 female, 7 males; 5 Deaf, 10 deaf, and 10 HoH; mean age=22.1, SD=3.1) from our university campus and surrounding area for an in-person interview. Our team included individuals with native fluency in both ASL and English, and thus we conducted the interviews, conversing with the participants using their preferred language (no need for interpreters) in a quiet room. Two members of our team were present (one acting as interviewer, the other as notetaker). Participants received $40 cash compensation for this 60-minute interview. After the demographic questions, we began the interview by asking participants how long they typically watch online videos (they averaged 102.6 minutes of daily video time) and which device they used the most to watch videos (56% *laptops*, 36% *smartphones*, and *televisions* tied with *desktop computers* for 4%). Then, the participants were asked to do three tasks (see Figure 3) related to the most/least important genres (a Latin-squares rotation for the task order was used - ABC, ACB, BCA, CBA):

TA) Answer the BWS questions using our survey UI on a laptop we provided.

TB) Look at cards (genres) and divide them into three groups: *Most important, Neutral, and Least important.*

TC) Look at cards (genres) and sort them into descending order of importance.

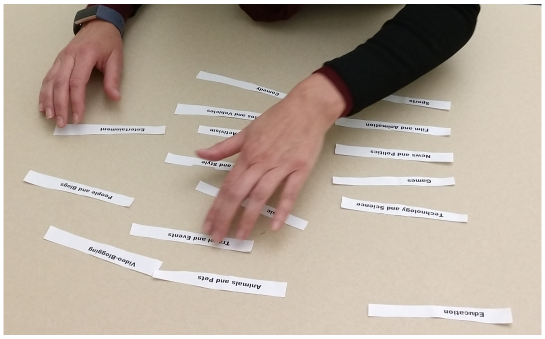


Figure 3. Participant ranking video genres via cards.

Table 1. Ranking results from BWS and Card-Sorting, with a Two-One-Sided-Test equivalence testing, indicating equivalence for all genres but one ("Tech and Science").

Online Video Genre	Rank from BWS	Rank from Card	p value from TOST
Animals & Pets	12.72	12.28	0.0165 *
Autos & Vehicles	10.72	10.48	0.0093 *
Comedy	9	8.64	0.0142 *
Education	2.6	2.2	0.0347 *
Entertainment	7.84	8.08	0.0113 *
Film & Anim	8.52	8.16	0.0139 *
Games	13.4	13.32	0.0071 *
HowTo & Style	8.72	9.48	0.0252 *
Music	11.32	11.44	0.0069 *
News & Politics	2.76	2.52	0.0128 *
NonProfits	6.16	6.44	0.0118 *
People & Blogs	9.04	9.64	0.0210 *
Sports	12.64	12	0.0276 *
Tech & Science	4.04	4.84	0.0514
Travel & Events	7.16	6.88	0.0113 *
Videoblogging	9.36	9.6	0.0092 *

5.2 Analysis of In-Person Interview Data

To evaluate whether BWS (**TA**) and card-sorting (**TC**) resulted in similar ranking results, we computed the mean ranking from both methods (Table 1). For each genre, we checked whether the rank from BWS was statistically equivalent to the rank from Card-Sorting, by calculating [31] the paired two one-sided t-test (TOST, using Welch's t-test which does not assume equal variances); p-values below 0.05 indicate statistical equivalence of the means. Overall, the ranks obtained from each of these methods were similar, with only one genre for which we did not measure statistical equivalence (*Tech and Science*).

5.3 Evaluating the BWS User-Interface

The 25 interview participants labelled (**TB**) and ranked (**TC**) the genres in 2.882 minutes on average (these two tasks were essentially done together as the participant arranged slips of paper on a table), and they completed the BWS questions (**TA**) in 2.83 minutes. This suggests that the time needed for each task was comparable.

Immediately after the BWS, participants were asked three questions about the usability of the experience:

- **Were the BWS directions clear and understandable?** (*No/Yes*). Almost all (24 of 25) thought BWS directions were clear and understandable.
- **I found the BWS UI to be confusing** (*5-point Likert*). Participants indicated that they did not find it confusing: (52% *Strongly Disagree*, 16% *Disagree*, 16% *Neither agree nor disagree*, and 16% *Agree*).
- **After completing the BWS, any opinion/feedback on the interface?** (*open-ended*). Most responses were positive (60%); participants did not indicate that it was hard to choose amongst the items, but it gave them a chance to think deeper about the genres: "*Paper better but like laptop too because its random and tricks you if you don't pick the same answers - you have to take it more seriously.*" (P7) Some participants mentioned that the test was *repetitive* as one participant signed it beautifully: "*Some feels like repeat. Trick question? Not sure which one answer...*" (P2)

5.4 Discussion

As no prior published research had ever deployed BWS among DHH individuals, we have distributed question prompts in English and ASL, and our validation analysis revealed that BWS obtained similar ranking results as in-person card sorting, and comparing the numeric ranks yielded **statistically equivalent** results. Almost all of our DHH participants were able to understand the BWS questions clearly and did not find the questions or the UI itself confusing. Thus, we conclude from our study that BWS can be a beneficial tool to rank a large set of items in a survey context with DHH participants; they were able to prioritize genres of videos using BWS. With this result, we could use BWS in a larger online survey (discussed below).

6 RQ2: Thresholds for Most/Least Important

While BWS is able to produce an overall ranking of a set of items, it can also be used to identify ordinal categorical labels, e.g. "*Important, Neutral, Not Important.*" However, to do this, it is necessary to identify threshold values for the BWS store for each of these ordinal levels, which can be used to partition items into these categorical groups. For this analysis, it was useful to consider the results of the "divide your cards into three levels" task from our in-person study (**TB**). We first needed to convert the numeric BWS scores (which are in the range of [-1, 1] with higher values corresponding to the best item) into categorical labels (*Important, Neutral*, and *Not important*) for each genre.[8] Our 25 participants labelled items as *Neutral* in task **TB** on 128 occasions. Of these scores, 75 of them were >0 and 53 of them were <0. The mean of the 75 BWS scores that were >0 was 0.1237, and the mean of the 53 BWS scores that were <0 was -0.3052. To define our thresholds, we used these scores to decide that the BWS range for the *Neutral* classification should be **[-0.3052, 0.1237]**. By this determination, we can then declare that any video genre that with a BWS score >0.1237 could be categorically labelled as "Important," and any video genre with a BWS score below -0.3053 could be labelled "Not Important." In the next section we will conduct BWS in a large online survey, and we use these thresholds in the subsequent analysis.[9]

7 RQ3: Most/Least Important Genres

To address RQ3, DHH users responded to an online survey using BWS to identify genres important for captioning.

7.1 Online Survey

The same user-interface for BWS was used for this online survey. Unlike our in-person study, the online survey did not include the card tasks (**T_B** and **T_C**). Our screening criteria included: (1) Are you over age 18? (2) Do you have hearing loss or identify as Deaf or Hard of Hearing? and (3) Do you use captions when viewing television or online videos? Based on email advertising among DHH community groups and national organizations, individuals interested in participating contacted a research assistant by email to obtain a survey code in order to access the online server. We compensated our participants by conducting a raffle for $250 Amazon gift cards with odds set at 1:100. Our team received 239 emails from interested parties, and 151 finished the survey for a completion rate of ~63%. Participants were from 27 U.S. states, and they self-identified as Deaf (80), deaf (23), and hard of hearing (48). There were 64 males and 87 females, with ages 18-88 (mean 39.9, std. dev. 19.6). Most (116) reported using ASL, and 35 did not. Their education levels varied: 8 high school, 45 some college, 52 bachelors, 38 graduate / masters, and 8 doctoral degrees. The participants completed the BWS questions in 3.81 minutes and 14.5 rounds on average, and we computed the BWS scores shown in Table 2 (ordered by score, with the highest score for the best item, and the best/worst columns displaying the count).

7.2 Results of BWS and Statistical Checks

Notably, when a BWS procedure is concluded, it is possible to produce a rank of the items in two ways: For each individual participant or across all participants. We are interested in this second type of ranking, since we are interested in the opinion across all the DHH respondents.

Table 2. Ranked BWS scores for the 16 online-video genres in our final large study, along with the count of the number of best/worst votes collected for each genre.

Online Video Genre	***BWS Score***	***#Best Votes***	***#Worst Votes***
News & Politics	0.679	396	11
Education	0.641	367	4
Tech & Science	0.436	269	22
Film & Anim	0.138	163	85
Entertainment	0.123	147	77
NonProfits	0.099	151	95
Comedy	0.074	126	84
Travel & Events	0.048	122	95
HowTo & Style	0.035	156	136
People & Blogs	-0.141	85	165
VideoBlogging	-0.146	75	158
Autos & Vehicles	-0.263	54	203
Music	-0.362	52	257
Sports	-0.374	58	270
Animals & Pets	-0.476	23	293
Games	-0.510	24	313

We calculated some statistics regarding the distribution of BWS scores in Table 2: [Shapiro-Wilk W=0.9374 with *p*=0.3179, Skewness=0.4377, Kurtosis=2.3482]. A Shapiro-Wilk normality test signified that the scores were normally distributed: positive skewness revealed that the participants agreed on the "best" choices more strongly than "worst" choices, and the excess kurtosis (platykurtic distribution) indicated that there were few outliers in the scores (not one genre stood out from the others). Table 2 also includes columns that show the total number of times that each genre received a "best" or a "worst" vote across the entire study (i.e. "most" or "least" important in this N-tuple). We checked the relationship between the best and worst votes, using Spearman's rank correlation coefficient: [ρ= -0.9198, *p*-value=4.622e-7], which indicated that participants' opinions did not conflict with each other often.

To evaluate how well the ranking opinion of each participant agreed with the ranking opinion of others, we calculated the Split-Half Reliability (SHR), which is better suited to evaluating the reliability of ranking judgments than many other inter-rater agreement statistics, as discussed in [25]. SHR repeatedly divides the dataset into random halves then calculates the correlation between the halves. We calculated[10] SHR and obtained a

[8] Provided by Svetlana Kiritchenko and Peter Turney at the NRCC: https://www.svkir.com/resources/Best-Worst-Scaling.zip, the Perl program is "get-scores-from-BWS-annotations-counting.pl".

[9] Our intention in listing these thresholds was for completeness; however, the specific threshold values likely depend upon the composition of the entire set of genres included in this set, due to the comparative nature of BWS. Future researchers are cautioned that if the set of genres under consideration were to differ, then the specific threshold values may differ.

[10] See the Perl program SHR-BWS.pl provided by S. Kiritchenko and P. Turney at: https://www.svkir.com/resources/Best-Worst-Scaling.zip.

Spearman correlation of 0.972 +/- 0.0113, which is a strong indicator of the reliability of the annotations in the dataset. Participants tended to agree on the ranking of the genres.

7.3 Discussion

As discussed in RQ2, it can also be useful to interpret BWS scores as categorical labels. Applying the threshold we calculated in RQ2 to the BWS scores in Table 2 allowed us to determine the final list of genres:

Most important[11] genres for captioning: News and Politics, Education, Technology and Science, Film and Animation, Entertainment.

Least important genres for captioning: Games, Animals and Pets, Sports, Music.

Overall, we found that participants generally agreed that several genres were most important (e.g. *News and Politics* and *Education*) and least important (e.g. *Sports* and *Animals and Pets*) to be captioned. Now that we had a clear idea of what genres are most and least important to be captioned for the DHH community, we are interested in figuring out **why** the genres were prioritized the way they were.

8 RQ4: Issues for Most/Least Important Genres

Now that we have discussed the quantitative results from our large online survey using the BWS methodology, we wanted to return our attention to some comments from participants in our earlier in-person study, which had included some open-ended interview questions:

- For the genre that is ranked #N, could you explain why it is the [most / least] important genre to be accurately captioned?
- For the genre that is ranked #N, could you give one example of when bad captioning for this genre [ruined / did not ruin] the experience for you?

We asked those questions for the best 4 and least 4 genres (as had been ranked by that individual participant), so there were 16 potential comments per participant (we filtered out "*no comment*" or equivalent replies). For our analysis of the responses, our annotation team was composed of a hearing faculty advisor with a specialty in HCI/accessibility and 4 DHH students (with native fluency in ASL) at the PhD, masters, and undergraduate level. The team followed the guidelines in [50] for two rounds of affinity diagramming and thematic coding to identify issues of interest to our participants and analyzed 256 comments (5,465 words).

8.1 Most Important Genres

For those genres our participants had personally considered most important to be accurately captioned, our team discovered these themes from 75 comments:

- **Society / World, (N=25)** Many participants wanted to know more about the world around them, e.g. "*A lot of hearing people know whats happening everyday but its not fair that deaf people are not as aware of what happening in the world and are behind in general and its very important for the deaf community to be up to date and to be aware of current events.*" (P6, *News and Politics*, ranked #2)
- **School / Education, (N=16)** Others wanted accurate captioning to support them as a student or for career preparation, e.g. "*My major is science. If video do not have captions I have to research longer to find the similar information. Long time to find right similar video.*" (P12, *Technology and Science*, ranked #3)
- **Self / Lifestyle, (N=14)** Some participants wanted to get information to apply to their own lives, e.g. "*I can learn how to fix the engine or something rather than going to the shop since it is cheaper for me to do it myself. I can know what exactly to do or what parts to order.*" (P13, *Autos and Vehicles*, ranked #3)
- **Updates / News, (N=9)** A few participants indicated that they wanted to keep up-to-date with information, e.g. "*Personally like science and technology, pictures are fine but need to know name and information.*" (P24, *Technology and Science*, ranked #3)

We analyzed how bad captioning impacted participants' experience and we found those themes from 65 comments:

- **Vocabulary Mix-Ups, (N=19)** The most problematic issue was specific words being incorrectly captioned, e.g. "*If I try to cook food or something if the video is not correctly captioned then I will mess up cooking.*" (P9, *How-To and Style*, ranked #2)
- **Hard to Comprehend, (N=8)** Other participants commented that the bad captioning caused them to increase their cognitive processing, e.g. "*Word choice causing misunderstandings. Run on sentences, hard to figure out what you mean and causes me to fall behind.*" (P21, *Technology and Science*, ranked #3)
- **Wrong Timing, (N=7)** Some participants were frustrated captions appeared at the wrong time: "*Example, debate. The people talk too fast and the captions can't catch up. Need to include the person's name. If no name, how do I know whos talking.*" (P10, *Education*, ranked #1)

8.2 Least Important Genres

Conversely, for those genres our participants did not personally consider important to be accurately captioned, there were 116 comments, and several themes emerged:

- **Information Already in the Video, (N=40)** Many participants expressed that the information they desired was already in the video in alternate form, e.g. "*They have score information. People use gestures to sign. I understand most of the visuals. I do not need captions.*" (P22, *Sports*, ranked #15) and "*Shooting games – the sounds aren't shown in the captions but*

[11] For the interested reader, Fleiss' Kappa for the most/least important genres: most=0.204 (*fair* agreement), least=0.0252 (*slight* agreement).

this doesn't really affect my experience." (P21, *Games*, ranked #16)

- **Video is Primarily Visual, (N=31)** Other participants mentioned that they wanted to see the visual information in the video thus the captioning wasn't important, e.g. "*I do not watch captions and I am not interested. I like to see cars.. I am visual.*" (P19, *Autos and Vehicles*, ranked #16) and "*If they have lousy captions, they do not bother me. I want to see visual and how do they make.*" (P18, *How-To and Style*, ranked #14)
- **Information Elsewhere Online, (N=14)** Some mentioned information available elsewhere: "*I'm deaf and I don't care about the words. I can hear vibrations... Can find words to lyrics online. (as opposed to news and politics, education, science and tech for example it is hard to find a script).*" (P5, *Music*, ranked #16) and "*Info not important, I can be patient and get from other place.*" (P2, *Technology and Science*, ranked #15)

8.3 Discussion

Analysis of these open-ended interview questions provided us with insights about the factors that affected DHH individual's prioritization of video genres. Their responses typically supported the results from **RQ3**: which genres were most or least important to be captioned. For example, there was a recurring theme in participants' comments where they wanted to be aware of what was happening in society and the world, which embodies the essence of the *News and Politics* genre (ranked one of the most important genres as shown in the results of **RQ3**). As another example, participants mentioned in their interview responses that they generally didn't need captions when there was a lot of visual information. The genres *Sports, Animals and Pets*, and *Games* were among the lowest ranked, and all are highly visual and thus captions aren't a requirement to fully understand the informational content presented in the video for those genres.

9 Conclusion, Limitations, and Future Work

In this project, the Best-Worst Scaling method was **effective** in gathering judgments in an online survey among a large number of DHH participants about which genres of online video they believed were important to be accurately captioned (Table 2). Participants in an initial in-person study enabled us to verify the efficacy of our BWS implementation (**RQ1**) and to identify thresholds of BWS values that categorically separate Important, Neutral, or Not Important judgements of DHH individuals (**RQ2**). With our contribution, researchers can now confidently survey DHH participants online to rank items, which may be useful for investigating many accessibility topics.

The data we gathered from participants in the online survey enabled us to calculate a final list of most-important and least-important genres (**RQ3**). Finally, a thematic analysis of comments from DHH interview participants revealed factors related to users' judgements about whether video genres were important for captioning (**RQ4**). The findings from our work will benefit researchers who are interested in how DHH viewers perceive the **importance of captions** for different online video genres, and it can help companies who provide online video content to prioritize how to improve accessibility of videos. Our findings also support future efforts to create datasets of online videos, for use in training or evaluating new technologies for captioning.

This study does have several limitations. While we have provided a prioritization of videos genres into broad categories in this work (categories used by a major video-sharing site), a subsequent study can prioritize sub-genres using our validated BWS methodology. While during our 25 in-person interview sessions, no participants expressed any confusion with what the title of each genre represented, if future researchers were to investigate more subtle sub-categories of video genres, it would be essential to ensure that participants understood what each genre title meant.

Another limitation was that our study included only participants from the United States – it would be valuable to determine whether DHH people in other parts of the world have different preferences; difference in culture could certainly result in different video interests. In addition, this study would have also benefitted from an even larger set of participants; such data could also allow us to analyze whether there are other demographic factors that may influence users' responses about genre priorities.

In addition to addressing these limitations above, we also foresee several avenues for future work: Although we had described previously that monitoring what people watch currently may not be a good proxy for what they want to watch or what would be important to have captioned (since current lack of captions may influence this), it would still be interesting to triangulate the results of our current study by analyzing DHH participants actual video viewing habits (by monitoring this over time or through an observational field study). Additionally, since the BWS algorithm worked well with DHH participants, we are interested whether BWS could be used to investigate how DHH people prioritize other needs, interests, or requirements. For instance, such an approach could be used to ask users about which types of environmental sounds (e.g. a doorbell) they might like to have access to, which would be valuable for researchers interested in sound-detection applications. Lastly, our original interest in video genres was motivated by an interest in the use of automatic speech recognition for captioning. In future work, we plan to create a stimuli collection consisting of the most/least important genres of online videos (with annotated transcripts) to provide a corpus as a resource for the research community for use in captioning studies with DHH individuals.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. 1763569; the Dept. of Health and Human Services under Award No. 90DPCP0002-01-00; a Microsoft Research Dissertation Grant, and a Microsoft AI for Accessibility (AI4A) Award.

REFERENCES

[1] John Albertini and Connie Mayer. 2011. Using Miscue Analysis to Assess Comprehension in Deaf College Readers. *The Journal of Deaf Studies and Deaf Education (JDSDE)* 16, 1 (2011), 35–46. DOI:http://dx.doi.org/10.1093/deafed/enq017

[2] Jon P. Barker, Ricard Marxer, Emmanuel Vincent, Shinji Watanabe. 2017. The CHiME challenges: Robust speech recognition in everyday environments. In: Watanabe S., Delcroix M., Metze F., Hershey J. (eds.), *New Era for Robust Speech Recognition*. Springer, Cham, 327-344. DOI:https://dx.doi.org/10.1007/978-3-319-64680-0_14

[3] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery (ACM), New York, NY, USA, 155–164. DOI:http://dx.doi.org/10.1145/3132525.3132541

[4] Denis Burnham, Jordi Robert-Ribes, and Ruth Ellison. 1998. Why captions have to be on time. In *Proceedings of the International Conference on Auditory-Visual Speech Processing* (Sydney, Australia) *(AVSP '98)*. ISCA, Baixas, France, 4.

[5] Juan Cao, Yong-Dong Zhang, Yi-Cheng Song, Zhi-Neng Chen, Xu Zhang, and Jin-Tao Li. 2009. MCG-WEBV: A benchmark dataset for web video analysis. *Beijing:Institute of Computing Technology* 10 (2009), 324–334.

[6] Jiyoung Cha. 2013. Does genre type influence choice of video platform? A study of college student use of internet and television for specific video genres. *Telematics and Informatics* 30, 2 (2013), 189 – 200. DOI:http://dx.doi.org/10.1016/j.tele.2012.09.003

[7] C. Chapdelaine, V. Gouaillier, M. Beaulieu, and L. Gagnon. 2007. Improving video captioning for deaf and hearing-impaired people based on eye movement and attention overload. *Proceedings of the SPIE* 6492 (2007), 11. DOI:http://dx.doi.org/10.1117/12.703344

[8] Eli Cohen. 2009. Applying best-worst scaling to wine marketing. *International journal of wine business research* 21, 1 (2009), 8–23.

[9] Michael Crabb, Rhianne Jones, Mike Armstrong, and Chris J. Hughes. 2015. Online News Videos: The UX of Subtitle Position. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery (ACM), New York, NY, USA, 215–222. DOI:http://dx.doi.org/10.1145/2700648.2809866

[10] Paul G. Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66 (2016), 4 – 19. DOI:http://dx.doi.org/10.1016/j.jesp.2015.07.006

[11] Deborah I Fels, Daniel G Lee, Carmen Branje, and Matthew Hornburg. 2005. Emotive captioning and access to television. *AMCIS 2005 Proceedings* (2005), 300.

[12] Adam Finn and Jordan J Louviere. 1992. Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing* 11, 2 (1992), 12–25.

[13] Terry N Flynn, Jordan J Louviere, Tim J Peters, and Joanna Coast. 2007. Best–worst scaling: what it can do for health care research and how to do it. *Journal of health economics* 26, 1 (2007), 171–189.

[14] Terry N. Flynn and A.A. J. Marley. 2014. *Best-Worst Scaling: Theory and Methods*. Edward Elgar Publishing, Cheltenham, UK, 178–201. DOI:http://dx.doi.org/10.4337/9781781003152

[15] Stephen Gulliver and George Ghinea. 2003. How level and type of deafness affect user perception of multimedia video clips. *Universal Access in the Information Society (UAIS)* 2, 4 (01 Nov 2003), 374–386. DOI:http://dx.doi.org/10.1007/s10209-003-0067-5

[16] Timothy J. Hazen. (2006). Automatic alignment and error correction of human generated transcripts for long speech recordings. *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing*, INTERSPEECH 2006 - ICSLP.

[17] Ellen S. Hibbard and Deb I. Fels. 2011. The Vlogging Phenomena: A Deaf Perspective. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (Dundee, Scotland, UK) *(ASSETS '11)*. Association for Computing Machinery (ACM), New York, NY, USA, 59–66. DOI:http://dx.doi.org/10.1145/2049536.2049549

[18] Yun Huang, Yifeng Huang, Na Xue, and Jeffrey P. Bigham. 2017. Leveraging Complementary Contributions of Different Workers for Efficient Crowdsourcing of Video Captions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4617-4626. DOI:https://doi.org/10.1145/3025453.3026032

[19] Carl Jensema, Ralph McCann, and Scott Ramsey. 1996. Closed-Captioned Television Presentation Speed and Vocabulary. *American Annals of the Deaf* 141, 4 (1996), 284–292. http://www.jstor.org/stable/44401017

[20] Carl J. Jensema, Ramalinga Sarma Danturthi, and Robert Burch. 2000a. Time Spent Viewing Captions on Television Programs. *American Annals of the Deaf* 145, 5 (2000), 464–468. http://www.jstor.org/stable/44393238

[21] Carl J Jensema, Sameh El Sharkawy, Ramalinga Sarma Danturthi, Robert Burch, and David Hsu. 2000b. Eye movement patterns of captioned television viewers. *American annals of the deaf* 145, 3 (2000), 275–285.

[22] A.K. Jones, D.L. Jones, G. Edwards-Jones, and P. Cross. 2013. Informing decision making in agricultural greenhouse gas mitigation policy: A Best–Worst Scaling survey of expert and farmer opinion in the sheep industry. *Environmental Science & Policy* 29 (2013), 46 – 56. DOI:http://dx.doi.org/10.1016/j.envsci.2013.02.003

[23] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the Usability of Automatically Generated Captions for People Who Are Deaf or Hard of Hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery (ACM), New York, NY, USA, 165–174. DOI:http://dx.doi.org/10.1145/3132525.3132542

[24] Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*. 465–470. DOI:http://dx.doi.org/10.18653/v1/P17-2074

[25] Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. San Diego, California.

[26] Thomas Ksiazek, Limor Peer, and Kevin Lessard. (2014). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media & Society*. 18. DOI:http://dx.doi.org/10.1177/1461444814545073

[27] Raja Kushalnagar and Kesavan Kushalnagar. 2018. SubtitleFormatter: Making Subtitles Easier to Read for Deaf and Hard of Hearing Viewers on Personal Devices. In *Computers Helping People with Special Needs*, Klaus Miesenberger and Georgios Kouroupetroglou (Eds.). Springer International Publishing, Cham, 211–219.

[28] Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2013. Captions Versus Transcripts for Online Video Content. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (Rio de Janeiro, Brazil) *(W4A '13)*. Association for Computing Machinery (ACM), New York, NY, USA, Article 32, 4 pages. DOI:http://dx.doi.org/10.1145/2461121.2461142

[29] Raja S. Kushalnagar, John J. Rivera, Warrance Yu, and Daniel S. Steed. 2014. AVD-LV: An Accessible Player for Captioned STEM Videos. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility* (Rochester, New York, USA) *(ASSETS '14)*. ACM, New York, NY, USA, 287–288. DOI:http://dx.doi.org/10.1145/2661334.2661353

[30] Paddy Ladd. 2003. *Understanding Deaf Culture: In Search of Deafhood*. Multilingual Matters, Bristol, UK.

[31] Daniel Lakens. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8(4), 355-362. DOI:http://dx.doi.org/10.1177/1948550617697177

[32] Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best–worst scaling: Theory, methods and applications*. Cambridge University Press, Cambridge, MA.

[33] Jordan J Louviere and George G Woodworth. 1990. Best-worst scaling: A model for largest difference judgments [Working Paper]. *Faculty of Business, University of Alberta* (1990).

[34] Henry B. Mann and Donald R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (03 1947), 50–60. DOI:http://dx.doi.org/10.1214/aoms/1177730491

[35] Marc Marschark, Harry G. Lang, and John A. Albertini. 2001. *Educating deaf students: From research to practice* (1 ed.). Oxford University Press, Oxford, UK.

[36] Ichiro Maruyama, Yoshiharu Abe, Eiji Sawamura, Tetsuo Mitsuhashi, Terumasa Ehara, and Katsuhiko Shirai. 1999. Cognitive experiments on timing lag for superimposing closed captions. In *Sixth European Conference on Speech*

Communication and Technology (EUROSPEECH'99). 575–578. https://www.isca-speech.org/archive/eurospeech_1999/e99_0575.html

[37] Alex Molassiotis, Richard Emsley, Darren Ashcroft, Ann Caress, Jackie Ellis, Richard Wagland, Chris D. Bailey, Jemma Haines, Mari Lloyd Williams, Paul Lorigan, Jaclyn Smith, Carol Tishelman, and Fiona Blackhall. 2012. Applying Best–Worst scaling methodology to establish delivery preferences of a symptom supportive care intervention in patients with lung cancer. *Lung Cancer* 77, 1 (2012), 199 – 204. DOI:http://dx.doi.org/10.1016/j.lungcan.2012.02.001

[38} M. Montagnuolo and A. Messina. 2007. Automatic Genre Classification of TV Programmes Using Gaussian Mixture Models and Neural Networks. In *Proceedings of the 18th International Workshop on Database and Expert Systems Applications* (Regensburg, Germany) *(DEXA)*. 99–103. DOI:http://dx.doi.org/10.1109/DEXA.2007.92

[39] Jacob Nyarko and Kwaku Oppong Asante. 2015. Social Exclusion of the Deaf in Corporate Television Advertising in Ghana: A Pilot Study. *Journal of Communication* 6, 2 (2015), 284–295. DOI: http://dx.doi.org/10.1080/0976691X.2015.11884874

[40] Hearing Loss Association of America. 2019. Hearing Loss Basics - How to tell if you have hearing loss. (2019). https://www.hearingloss.org/hearing-help/hearing-lossbasics/

[41} Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental. Social Psychology* 45, 4 (2009), 867 –872. DOI:http://dx.doi.org/10.1016/j.jesp.2009.03.009

[42] Seth Ovadia. 2004. Ratings and rankings: reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology* 7, 5 (2004), 403–414. DOI:http://dx.doi.org/10.1080/1364557032000081654

[43] Susan J. Parault and Heather M. Williams. 2010. Reading Motivation, Reading Amount, and Text Comprehension in Deaf and Hearing Adults. *The Journal of Deaf Studies and Deaf Education (JDSDE)* 15, 2 (2010), 120–135. DOI:http://dx.doi.org/10.1093/deafed/enp031

[44] Dimitris Potoglou, Peter Burge, Terry Flynn, Ann Netten, Juliette Malley, Julien Forder, and John E. Brazier. 2011. Best–worst scaling vs. discrete choice experiments: An empirical comparison using social care data. *Social Science & Medicine* 72, 10 (2011), 1717 – 1727. DOI:http://dx.doi.org/10.1016/j.socscimed.2011.03.027

[45] Soraia Silva Prietch, Napoliana Silva de Souza, and Lucia Villela Leite Filgueiras. 2015. Application Requirements for Deaf Students to Use in Inclusive Classrooms. In *Proceedings of the 7th Latin American Conference on Human Computer Interaction* (Córdoba, Argentina) *(CLIHC '15)*. Association for Computing Machinery (ACM), New York, NY, USA, Article 5, 8 pages.DOI:http://dx.doi.org/10.1145/2824893.2824898

[46] Anni Rander and Peter Olaf Looms. 2010. The Accessibility of Television News with Live Subtitling on Digital Television. In *Proceedings of the 8th European Conference on Interactive TV and Video* (Tampere, Finland) *(EuroITV '10)*. ACM, New York, NY, USA, 155–160. DOI:http://dx.doi.org/10.1145/1809777.1809809

[47] Jonathan Rubin, Ryan Leisinger, and Gary Morin. 2014. 508 Accessible Videos-Why (and How) to Make Them. (June 2014). Retrieved April 29, 2019 from https:/digital.gov/2014/06/30/508-accessible-videos-why-and-how-to-make-them/

[48] Hillary M Sackett, Robert Shupp, and Glynn Tonsor. 2013. Consumer perceptions of sustainable farming practices: A Best-Worst scenario. *Agricultural and Resource Economics Review* 42, 2 (2013), 275–290.

[49] Brent N. Shiver and Rosalee J. Wolfe. 2015. Evaluating Alternatives for Better Deaf Accessibility to Selected Web-Based Multimedia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery (ACM), New York, NY, USA, 231–238. DOI:http://dx.doi.org/10.1145/2700648.2809857

[50] Anselm Strauss and Juliet M. Corbin. 1998. *Basics of qualitative research: Grounded theory procedures and techniques* (2 ed.). SAGE Publications, Inc., Thousand Oaks, CA, USA.

[51] Ba Tu Truong and C. Dorai. 2000. Automatic genre identification for content-based video categorization. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR)*, Vol. 4. 230–233. DOI:http://dx.doi.org/10.1109/ICPR.2000.902901

[52] Máté Akos Tündik, György Szaszák, Gábor Gosztolya, and András Beke. 2018. User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning. *Proceedings of Interspeech 2018* (2018), 2628–2632.

[53] Lucia Vesnic-Alujevic & Sofie VanBauwel (2014) YouTube: A Political Advertising Tool? A Case Study of the Use of YouTube in the Campaign for the European Parliament Elections, Journal of Political Marketing, 13:3, 195-212, DOI:http://dx.doi.org/10.1080/15377857.2014.929886

[54] J. Wu and M. Worring. 2012. Efficient Genre-Specific Semantic Video Indexing. *IEEE Transactions on Multimedia* 14, 2 (April 2012), 291–302. DOI:http://dx.doi.org/10.1109/TMM.2011.2174969

[55] Georgios N. Yannakakis and John Hallam. 2011. Ranking vs. Preference: A Comparative Study of Self-reporting. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction* (Memphis, Tennessee, USA) *(ACII '11)*, Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin (Eds.). Springer, Berlin, Heidelberg, 437–446. DOI:http://dx.doi.org/10.1007/978-3-642-24600-5_47

[56] Georgios N. Yannakakis and Héctor P. Martínez. 2015. Ratings are Overrated! *Frontiers in ICT* 2 (2015), 13. DOI:http://dx.doi.org/10.3389/fict.2015.00013

[57] Norman Bradburn, Seymour Sudman, Brian Wansink. 2004. *Asking questions: The definitive guide to questionnaire design--For market research, political polls, and social and health questionnaires* (Rev. ed.). San Francisco, CA, US: Jossey-Bass.

[58] Sofia Enamorado. 2019. Final CVAA and FCC Online Video Closed Captioning Rules. Retrieved on September 19, 2019, from https://www.3playmedia.com/2018/11/14/final-cvaa-and-fcc-online-video-closed-captioning-rules/

[59] Larwan Berke, Sushant Kafle, and Matt Huenerfauth. 2018. Methods for Evaluation of Imperfect Captioning Tools by Deaf or Hard-of-Hearing Users at Different Reading Literacy Levels. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Paper 91, 12 pages. DOI: https://doi.org/10.1145/3173574.3173665