# MEASURING THE EFFECTS OF BUFFERING AND INTERRUPTS ON USER EXPERIENCE FOR YOUTUBE AND NETFLIX

An Interactive Qualifying Project Report
submitted to the Faculty of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Bachelor of Science
By

Kevin MacDougall

George Randel

Advisor  Mark Claypool

Date: March 6, 2015

# Abstract

The quality of a streamed video has many influences, of which two are the buffer time, and the number of interrupts. There is no definitive research on what is the best buffer to interrupt ratio for a given content. The focus of research for this IQP is the effects of buffering and interrupts on user quality of experience for online video, with a specific focus on Netflix and YouTube. This IQP attempts the answer the question of whether or not viewers prefer YouTube videos to have short initial buffers and Netflix videos to have minimal interruptions. This is achieved using online surveys with test videos of varying buffer times and number of interrupts. The survey resulted in over 250 responses from users of Amazon’s Mechanical Turk. The results show there is no significant difference between the buffer time, and the number of interrupts with regards to user quality of experience.

# Contents

# Table of Figures

# 1 Introduction

Online video streaming viewership and advertising rates have grown rapidly [1,2]. There are many video streaming services available on the Internet for users to choose from and two of the most popular of which are YouTube and Netflix.  These two video providers have an interest in gauging a user's quality of experience (QoE) while viewing streamed video in order to adapt their services to better suit the user's preferences.

Video content provided by either YouTube or Netflix must be accessed via the Internet. Issues like network limitations and server load introduce the problem of variable data rates, which leads to waiting time [3]. When streaming video over the Internet, video providers cope with waiting time through initial buffer time and interrupts. The phenomenon of buffering is the downloading and storage of video information in a reservoir at one rate and the delivery of that information to the video player at another rate. The initial buffer time of a video is the time it takes to download a portion of video from a server in preparation of video playback. The downloaded information is stored in a buffer to be accessed later during playback [4]. This initial buffer allows the video playback to be more resilient to Internet conditions and lessens or eliminates the impact of varying data rates.  After a certain initial buffer time the video player starts taking information out of the buffer at the intended playback rate, called video playback. Video playback continues until all of the video is shown or until the buffer is empty. When the buffer is empty and new content has not arrived in time, the video playback stops, called an interrupt. During an interrupt, the buffer needs to be refilled before the video playback can continue. An example video playback on a client's machine can include ten seconds of initial buffer, video playback for two minutes, one interruption, and video playback until the end of the video.

The number of interrupts and the initial buffer time are not independent quantities. There is an inverse relationship between the two aspects of waiting time that force video providers to find a balance between initial buffer time and interrupt quantity. This inverse relationship means that a longer initial buffer will provide more protection against variable data rates and therefore fewer interrupts in video playback. Alternatively, a shorter initial buffer leaves the rest of the playback vulnerable to variable data rates, therefore causing the buffer to empty and interrupts to occur.

The research done in this IQP focused on video streaming from the two of the largest video providers, YouTube and Netflix. The goal is to test the hypotheses that users prefer YouTube videos to have short initial buffers and that users prefer Netflix videos to have minimal interrupts. Because of the tradeoff mentioned earlier, this also implies that users can tolerate more interrupts in a YouTube video and that users can tolerate a longer initial buffer in a Netflix video.

Users in this study viewed sample videos that are both representative of YouTube and Netflix and have been modified to intentionally add initial buffer and interrupt artifacts. The users were asked to provide qualitative and quantitative feedback on the perceived quality of the videos they watched in the form of an Internet based survey. This qualitative information was taken and compared to the quantitative data provided by the videos themselves with regard to their buffer times and interrupts.  The survey resulted in over 250 responses, of which 147 were deemed valid responses. The analysis of the data found that the hypothesis was not supported.

Chapter 2 reviews related work on the subject of video streaming and QoE data gathering.  It provides insight into video streaming infrastructure as well as coverage of existing research methods, metrics, and results.  Chapter 3 provides a description of the methods used for data gathering within this research along with the reasoning behind the decisions that were made. Chapter 4 outlines the applicability and usefulness of Amazon's Mechanical Turk in gathering large amounts of data from a diverse participant pool.  Chapter 5 examines the results of the survey and shows that the hypotheses cannot be supported through the analysis of the data.  Chapter 6 summarizes the results of the research and Chapter 7 suggests future work to be done to both improve the methodology and gather higher quality data.

# 2 Related Work

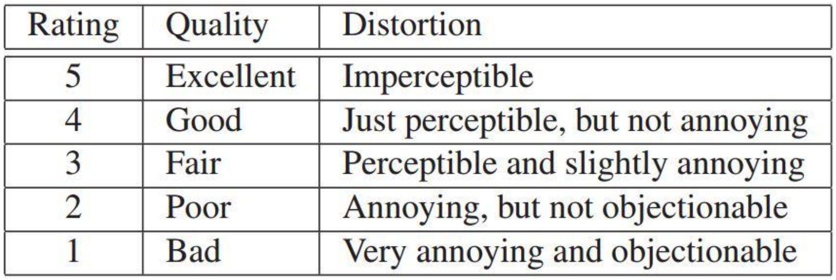
Despite the fact that the proliferation of online video streaming is relatively recent, there are a number of papers written about the importance of the user's experience while viewing streamed media. Useful information gathered from the related work pertains to user feedback metrics, feedback acquisition techniques, and existing user experience data. The few resources that do provide existing user experience data are instrumental in the development of a proper methodology for gathering and interpreting user feedback data. These researchers have used techniques like the Mean Opinion Score (MOS) or client side programs to gather the necessary information to make conclusions regarding the user's QoE, whereas others have looked into the effects that things like viewing environment and waiting time have on the user's QoE.

## 2.1 Mean Opinion Score (MOS)

User consumption of streamed media occurs mostly on personal computers and TVs in a comfortable environment [5]. It was important to develop a data gathering technique that does not remove the user from their preferred viewing environment in order to gather the most accurate data. Mean Opinion Score is a proven method for gathering qualitative feedback [6].

A Mean Opinion Score (MOS) is officially described as the subjective quality used to evaluate signal processing methods[7]. This standardized metric, though mostly used in the analysis of audio transmission, can be applied to the analysis of user experience while watching streamed video due to the fact that the rating mechanism is loosely dependent on the content being rated. The only consideration when applying

MOS to video quality versus audio quality is that the content is inherently different, which means that the analysis of the data must be handled accordingly.  MOS involves the rating of a video's quality on a 1 to 5 scale as seen in **Figure 1**.



| Rating | Quality | Distortion |
|---|---|---|
| 5 | Excellent | Imperceptible |
| 4 | Good | Just perceptible, but not annoying |
| 3 | Fair | Perceptible and slightly annoying |
| 2 | Poor | Annoying, but not objectionable |
| 1 | Bad | Very annoying and objectionable |

***Figure 1:*** *Typical MOS Scores [7]*

Some researchers [8] suggest that using MOS to manage the user's quality of experience can be potentially flawed because of the varying levels of predictability in human subjects.  The 5 levels of quality may have different meanings to different people, which can introduce error in data gathering.  A large sample size is required for statistically significant results.

## 2.2 Client Side Data Acquisition

In their research, Dobrian et. al. [9] describe the use of a client based application to collect raw data from the user's video player as he or she streams video from the Internet.  The two major accomplishments are the creation of the application that allows the collection of the data and the development of an analysis technique to expose correlations within the data.  The client side application works by having content providing partners embed the application within their players so when a user views a video on a participating site, the video playback data is sent to the researchers.  The playback data includes network statistics but sends back a video profile, which outlines buffer times and interrupts within a video. The application is capable of tracking data on a per view basis as well as on a per user basis. The per view scope focuses on the raw data from the videos themselves whereas the per user scope looks at how many videos a user views consecutively and attempts to correlate the number of videos watched to specific video imperfections like initial buffer time or interrupt quantity.

Despite finding that quantitative data gathering techniques when applied to a truly qualitative problem must “be used with caution and with a judicious appreciation of the context in which they are applied,” the researchers were able to gather some valuable insights into the correlation between video quality and user experience. Buffering Ratio, which is described as fraction of total session time spent in buffering, is by far the most important factor in determining the user’s experience across all video content types. The more time spent buffering, either at the beginning of the video or in the form of interrupts, the less engaged the user is. User engagement is defined as how likely one is to continue viewing a video or to continue watching subsequent videos. Initial buffer time also plays a large role in determining user engagement though less than buffering ratio. When it comes to live content, the download rate is especially important although the focus of this paper is limited to prerecorded media that is already encoded. The work done in this paper gives a comprehensive view on the correlation of qualitative video statistics to quantitative user experience scores.

## 2.3 Mobile vs. Computer Viewing Experience

Finamore et. al. [10] analyze the similarities and differences between watching a YouTube video on a mobile device verses on a desktop or laptop computer. The paper mostly finds that the user interface is almost identical no matter what device one views the video from; what differs, though, is the method used to buffer the video is less efficient on mobile devices than it is on computers. They found that unnecessarily large buffers were being downloaded to the phones and that users would frequently stop watching the video midstream, causing the buffer to be wasted. The part of this research that is relevant to this paper, however, is the analysis of the user's behaviors while watching YouTube videos. The paper reveals that about 40% of users terminate a YouTube video before playback has finished, which implies that the content of YouTube videos in particular plays an influential role in gathering data on user experience. It is suggested that YouTube should adopt a more conservative buffering scheme to minimize the investment on a per view basis because of the fact that users leaving the videos early. The information presented on the behavior of YouTube viewers is important to consider in the analysis of quantitative data gathered about YouTube video viewing statistics.

## 2.4 Waiting Time vs. QOE

Research done by Pessemier et. al. [11] suggests that there is a correlation between a user's subjective quality assessment of video playback and the objectively measured quantitative data of the viewing session.  This means that when interrupts and buffering times were added to the videos being watched by the users it was generally reflected in a more negative feedback score.  The experiment required users to view 14 different videos with varying encoding qualities, subject matters, length of initial buffers, and interrupt quantities on a mobile device while in a laboratory environment.  The study found that the encoding quality had significantly less of an impact on user feedback than the number of interrupts in the viewing session.

Specifically, a video with a waiting time of less than 20 seconds for the initial buffer was marked as acceptable by 75% of users in their post video surveys. Alternatively, 75% of users marked the video unacceptable if the initial buffer went past 60 seconds. The graph presented by the researchers in **Figure 2** shows the correlation between the qualitative information (Probability that the quality is not acceptable) and the quantitative information (Waiting Time) gathered in the study. This study is only considering the case where the user is viewing streamed media on a mobile device in a laboratory setting so the data is not necessarily representative of users on computers in their home but supports the concept of using a quantitative measure to predict a qualitative response and vice versa.

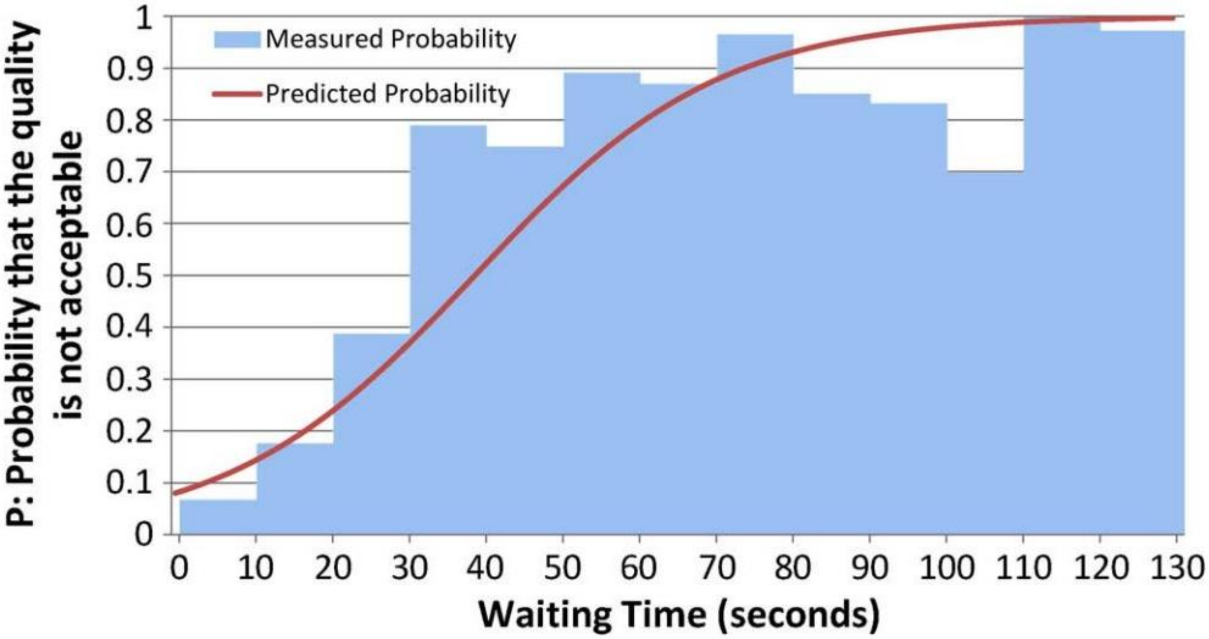


***Figure 2:*** *Correlation Between Waiting Time and the Probability that the Quality Is Not Acceptable [11]*

## 2.5 Network Infrastructure

Netflix manages to rely on very little network infrastructure of its own despite the fact that it consumes 31% of Internet traffic [12]. Netflix manages to provide content at such a large scale by the use of multiple Content Distribution Networks (CDNs) that provide constant streaming sources for the Netflix client to choose from [13]. Some of the only infrastructure that Netflix handles on their own is the servers to run user account registration and payment information. When a user requests a video, the data is first sent to Amazon cloud servers that are responsible for logging data, digital rights management and CDN routing. The source being used to stream the video supports Dynamic Adaptive Streaming over HTTP (DASH), which consists of videos that are encoded at different quality levels and split into small chunks that can be requested by the client. The quality level of the next chunk is calculated based on the download rate of the previous chunk. If video playback degrades then it is up to the host server to decide whether to downgrade quality or to look for a better suited CDN to provide the video content. A breakdown of Netflix's architecture can be seen in **Figure 3**.

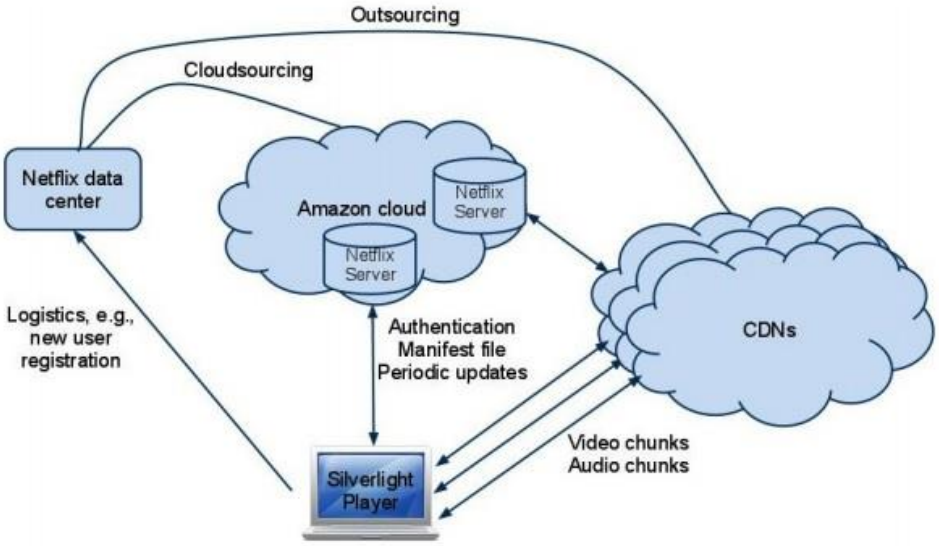


***Figure 3:*** *Netflix Video Streaming Architecture [13]*

YouTube uses a similar method of video streaming as Netflix by encoding video into several different qualities to be able to send the most appropriate chunk of video to the user based on current network conditions.  YouTube also relies on the use of CDNs to distribute the content for their video [14]. Of that, nearly 80% of interrupted playbacks are hypothesized to be the result of user termination due to the fact that the playback rate was well below that of the buffer rate of the video being watched [15].  It is logical then for YouTube to minimize its investment in providing long buffers for users who might not be interested in viewing the whole video, which is why a great deal of effort goes into providing streamlined video hosting service.

# 3 Methodology

To gather data on video quality for buffer to interrupt ratio for YouTube and Netflix, an online based user survey was used. The use of such a survey versus an in lab study allowed for a larger user base and removed the limitation of sampling people within the vicinity of WPI. One set of generic feedback questions was made for feedback on the YouTube and Netflix videos. The user feedback section consists of questions on how the user rates the video quality, and how the user rates his or her interest in the video. To identify the two formats, the classifications of professionally produced content versus amateur content, and long form versus short form content was used to categorize videos as being "Netflix" videos or "YouTube" videos.

## 3.1 Selecting Videos

YouTube and Netflix videos that are a representative samples were selected. This means videos were selected on the basis that they would mostly likely be associated with the platform under test. For example, in the case of YouTube, a representative video is one that is to the average viewer considered a homemade video and for Netflix, one that to the average viewer is seen as a professional video that is produced by a studio. The following two sections detail how the videos were selected for use in the study.

### 3.1.1 YouTube

There is a wide range of content on YouTube but there is far more amateur content on the site than professional content. For this reason a representative YouTube video was defined as amateur content, meaning not produced professionally or by a studio. A representative YouTube video requires little attention or effort is required by the audience to be entertained. An example of this is when a content creator films a cat doing something amusing or cute. Two videos like this were used in testing by taking the videos from YouTube and editing interrupts to replicate different buffering times and number of interrupts. The two videos were edited to be approximately one minute and 30 seconds, with 1 minute of content, and 30 seconds of buffer, or interrupts.

### 3.1.2 Netflix

Netflix only contains videos produced by studios or produced by professional filmmakers. For this reason a Netflix video is defined as professionally produced content. The long form content available on Netflix is generally well over 15 minutes, due to it mostly being TV shows and movies, which poses a challenge in having users in the study view Netflix content. It is unrealistic to have a user watch such long videos as part of a survey, because people will most likely not want to spend that much time. To solve this, movie trailers were used to represent Netflix content. The use of movie trailers make it clear to the survey taker that the video is representative of professional content, which is important in the differentiation of a Netflix video from a YouTube video. Netflix branding was placed in the video as well to further assure that the user thinks the content is representative of the videos provided by Netflix. This was done during the buffer times and during the interrupts.

## 3.2 Video Buffering and Interrupt Combining Method

The videos used in the study were organized based on the length of the initial buffer and the quantity of interrupts throughout playback. This buffering versus interrupt is based on that assumption that the initial buffer of the video is the independent variable and that the number of interrupts is the dependent variable. It follows that when the initial buffer is short there are many interrupts and as the initial buffer gets longer than the number of interrupts decreases. **Figure 4** shows an example curve of the relationship between these two factors. It is split into 5 different sections along the x-axis, which allows a score of 1-5 to be created with 1 being a video with the smallest initial buffer and most interrupts and 5 being a video with the largest initial buffer and fewest interrupts. This scoring system of 1-5 allows the videos used in this study to be scored quantitatively with a universal metric, that is consistent across questions.

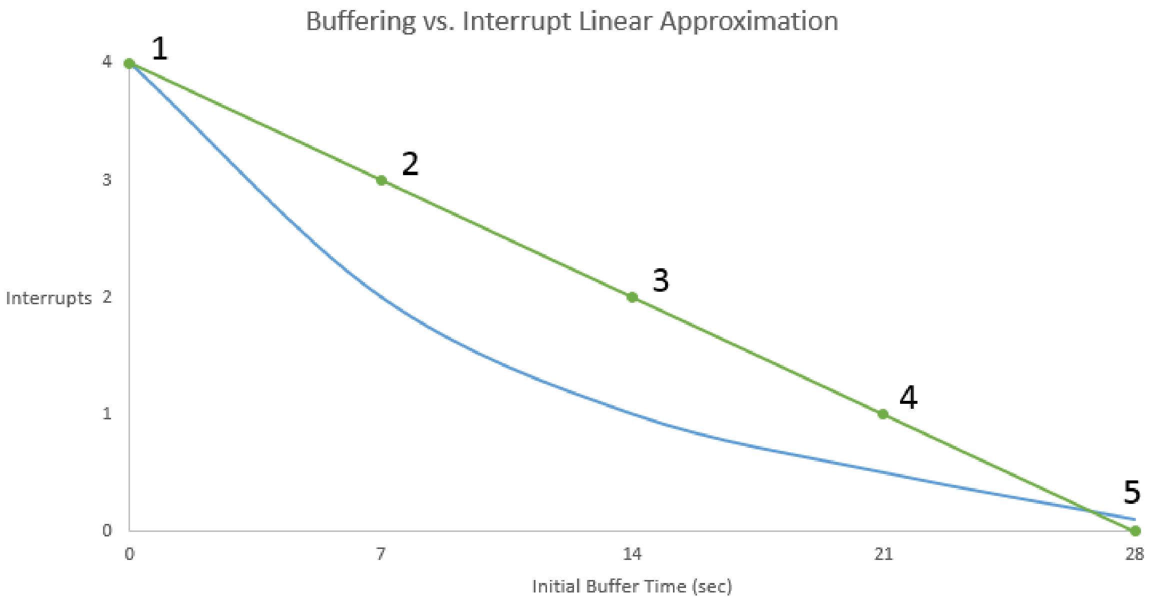


***Figure 4:*** *Interrupts vs. Initial Buffer Time Combining Method*

### 3.3 Creation of Test Videos

The videos that were selected to represent the two different types were then edited to introduce buffering and interrupt screens that are analogous to what would be seen on the platform under test. The videos were be encoded in such a way that the vast majority of local video players, like Windows Media Player, QuickTime, and VLC, can play them to allow for the most compatibility.

The videos were edited using Adobe Premier and encoded using H.264. This means that in the case of YouTube a series of spinning circles were played during the initial buffer, as can be seen in **Figure 5**, and during an interrupt an overlay of spinning circles was placed on the paused frame as seen in **Figure 6**. For Netflix, the only buffering screen video that could be found was from the Xbox, but it is clear that it is a buffering screen from Netflix. A similar method was used to edit the movie trailers for the Netflix test videos with an initial buffer screen and animated overlay during the interrupts. The initial buffering times and number of interrupts were added based on the previously discussed rating system, thus creating 10 variants for each video.

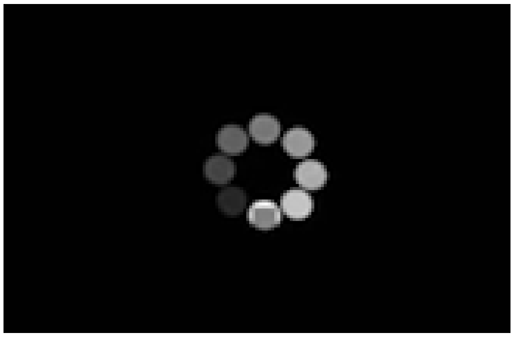


***Figure 5:*** *YouTube Initial Buffer Screen*

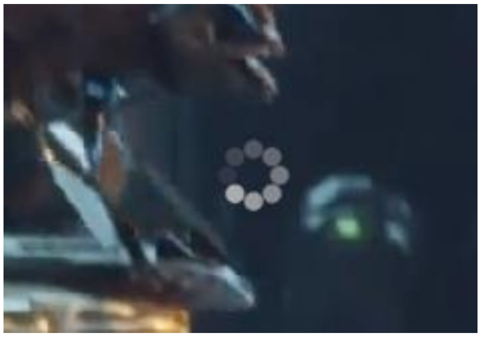


***Figure 6:*** *YouTube Interrupt Screen*

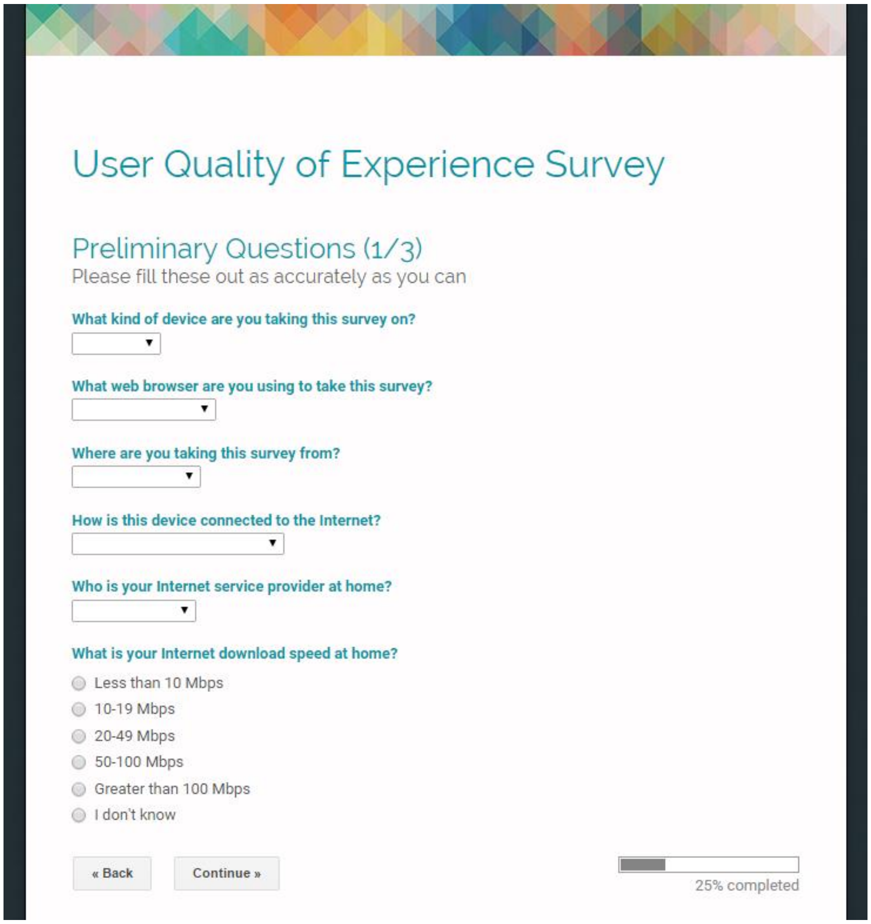
For the survey these videos were downloaded to the user's computer to avoid the possibility of additional buffering times or interrupts that might occurring if the videos were streamed to the user. To ensure that most users can play the videos, the videos were encoded in such a way to allow for the most compatibility. The most supported codec in use is H.264. Both Windows and Mac default players support H.264 and most Linux players will support it as well.

## 3.4 Survey

To attempt to gather the largest user base an online Internet based survey was conducted. While an in person laboratory study would likely produce more in depth responses from the users, a larger study can test a broader demographic and gather more data for statistical significance [16]. The following sections detail how the survey was built and recorded.

### 3.4.1 Survey format

The survey was built using Googles survey engine. Google's survey engine is free and exports the user's responses to a Google spreadsheet, which can be easily moved to Excel. Along with Google's survey engine, Mechanical Turk was used to gather responses. This will be discussed in section 4. A page from the survey can be seen in **Figure 7**.



User Quality of Experience Survey

Preliminary Questions (1/3)

Please fill these out as accurately as you can

What kind of device are you taking this survey on?

What web browser are you using to take this survey?

Where are you taking this survey from?

How is this device connected to the Internet?

Who is your Internet service provider at home?

What is your Internet download speed at home?

○ Less than 10 Mbps
○ 10-19 Mbps
○ 20-49 Mbps
○ 50-100 Mbps
○ Greater than 100 Mbps
○ I don't know

***Figure 7:*** *Google Forms: Survey Page*

The next part of the survey has the videos that were played to the users. The survey provides instructions to the user on how to download the video, and the download site created, serves a randomly selected video. This gives each user one of the five videos at random.

### 3.4.2 Survey Questions

The final consideration is what questions to ask the users. The user's response needs to be a good representation of what he or she thinks of the different buffering and interrupt conditions presented in the videos they watch. An initial questionnaire prompts the users to answer some simple questions about his or her Internet video viewing habits and relevant information such as if the users is using a wired or wireless connection and what device he or she is taking the survey on, as well as specifics on how much he or she watches Netflix or YouTube, and what his or her rating of the given service is. The users then download and view a test video that is representative of Netflix content. The user is prompted to provide feedback on his or her QoE after watching the video based on the 1-5 MOS scale. This process of downloading videos, viewing them, and providing QoE feedback is done four times in total; twice for Netflix sample videos and twice for YouTube sample videos.

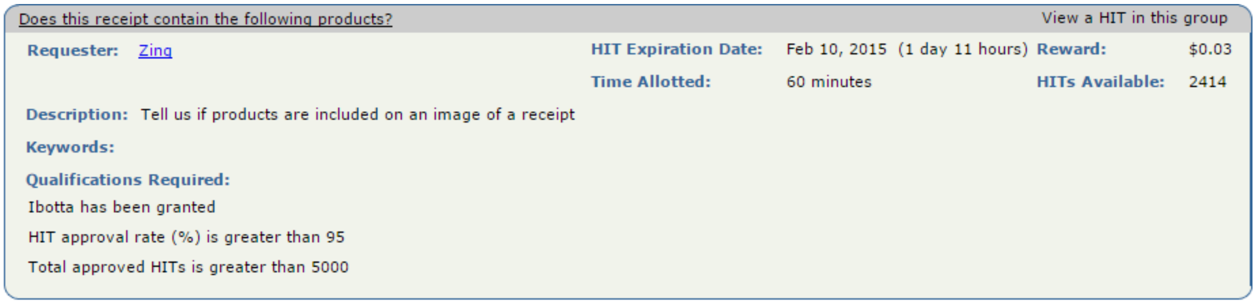
# 4 Mechanical Turk Distribution Platform

The problem of getting the survey out to a large and diverse pool of participants is addressed by using Amazon's Mechanical Turk (MTurk). Mturk is used for a large human workforce that can perform certain tasks that computers are currently unable to do. This is the motivating reason behind Amazon's investment in MTurk and the reason why we chose it as the survey distribution platform. MTurk is a free to join online tool that allows a requester to post a job for workers to complete for a small monetary reward. In our case, a requester account was set up for the purposes of distributing the Google Forms survey to the desired participant pool. The participant pool at MTurk is only limited to those who have access to the Internet and the time to complete at least one of the hundreds of thousands of available jobs at any moment. MTurk users range from the casual browser who completes one or two jobs to a more serious MTurk worker who consistently completes jobs. The requester can set qualifications on who can complete their jobs. For instance a requester can set a criteria that only workers who have passed the required tests to gain the necessary qualifications are eligible to complete a job. The requirements were set low in order to attract the most number of workers. Specifically, workers were only required to have an approval rating greater than or equal to 90%. An approval rating is a reflection of how well a specific worker has performed on other jobs, or Human Intelligence Tasks (HIT) as MTurk calls them.

## 4.1 Human Intelligence Tasks (HIT)

Computers are excellent at providing raw qualitative data for information such as the user's download rate, interrupt quantities, and initial buffer length but they fall short when it comes to interpreting those factors into a meaningful QoE measurement. HIT is a question that needs answering by a human because it cannot be answered by a computer. Posting a HIT allows for people to take the survey and provide their personal feedback on their QoE.

All the information about a HIT is included within its listing, which is visible by a worker. The listing contains the title and description of the job, expiration date, time allotted, number of HITs available, and the reward value. The title and description provides the bulk of the information about what the worker can expect to do if he or she decides to accept that HIT and the required qualifications of the worker set by the requester. The expiration date shows how long a listing is active. After that date the HIT is no longer active and no worker can accept it. The time allotted tells the worker how much time he or she has to work on a HIT once it is accepted. If the time allotted is exceeded by the worker, then his or her response is rejected and the HIT becomes available again. The number of HITs available shows how many more copies of the same HIT are available for workers to accept. As soon as a worker accepts a HIT, one less HIT is available for all other workers. MTurk ensures that one worker is only allowed to accept a HIT one time. This means that once a worker accepts a HIT, he or she either submits it or is penalized for going over the time limit that has been set by the requester. This penalty negatively impacts the worker's account reputation, signaling a poor worker. Finally, the listing gives the reward value for the worker if the task is completed. An example HIT listing can be seen in **Figure 8**.



Does this receipt contain the following products? View a HIT in this group

Requester: Zing

HIT Expiration Date: Feb 10, 2015 (1 day 11 hours)

Reward: $0.03

Time Allotted: 60 minutes

HITs Available: 2414

Description: Tell us if products are included on an image of a receipt

Keywords:

Qualifications Required:

Ibotta has been granted

HIT approval rate (%) is greater than 95

Total approved HITs is greater than 5000

***Figure 8:*** *Example Listing Showing Title and Description of a HIT*

## 4.2 Monetary Incentives

The driving factor for workers to complete HITs is the monetary reward they receive. The requesters attach a monetary value to each HIT in order to pay the worker for completing it and they also have to pay a 10% service charge to MTurk for each HIT they post.

A worker can claim the reward listed in the HIT if the requester approves his or her submission.  This means that the requester must review the worker’s submission either manually or automatically before any money is transferred to the worker.  The review process involves the requester checking to see if the worker’s submission meets the qualifications set by the requester.  Once the submission has been approved, the reward amount is credited to the worker’s account. Reward money is accessible by the worker through an Amazon payment account which can be tied to a bank account for direct deposit.

For the requester, the process starts with either an Amazon payments account, a credit card, or a bank account.  An established Amazon payments account or a credit card allows the requester to transfer money upon creating a requester account, whereas setting up a bank account requires at least a week to transfer money.  This money is put into what is called a HIT prepaid balance that is tied to the requester’s

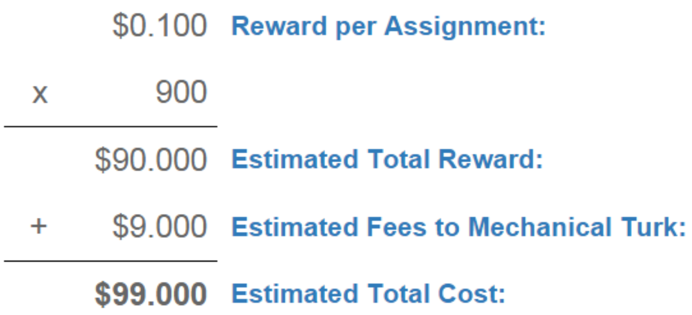
MTurk account. This balance can be used to fund the requester’s HITs. Once a portion of the prepaid balance is allocated to a certain HIT it cannot be touched by the requester until the portion is used up by paying workers, the HIT expires, or the requester cancels the HIT, at which point the money is transferred back to the prepaid balance. While the HIT is active, the reward amount along with MTurk’s 10% is transferred out of the money allocated to the HIT every time a submission is accepted by the requester. If there is still money left in the prepaid balance at the end of the HIT, the requester can then transfer the money back to the original purchase option and leave nothing in their requester account.

## 4.3 Posting and Managing a Job

The process of posting and maintaining a HIT on MTurk is done with a requester account. Such an account can be made by using an existing Amazon account or by creating a new account just for this purpose. Once the account is set up there are four main tasks involved in posting and maintaining a HIT. The first task is designing the content of the HIT. Second is deciding the parameters of the listing, like allotted time and reward amount. Third is monitoring the HIT while it is active and accepting or rejecting responses. Last is closing out the HIT when desired or when it expires.

The content of the HIT is the task that the worker will be completing. MTurk provides many options when it comes to designing the layout of the task. There are templates for multiple-choice answer, picture identification, etc. that can be easily modified. The template used in this research is one that allows the worker to access a link to an external survey site and then enter in a unique code upon completion of the survey to ensure that the worker has actually completed the external task.  Once the layout of the task is complete, the requester can move on to publishing it.

The information in the listing needs to be filled out, similar to **Figure 8**, in order to make the HIT available to the workers.  The title and description should provide an accurate representation of what is required of the worker if he or she decides to accept the HIT.  The requester also has the choice to set the worker qualification requirements to allow only the desired workers to access the HIT. For instance, a worker can be qualified as an expert in a certain task or he or she can have a high HIT acceptance rate. In order to increase the number of responses for this research, the worker qualification requirements were set low so that more users would be able to access the HIT. Setting the expiration date and time limit defines when the workers have time to accept and work on the HIT.  In another effort to increase responses, the expiration date for the HIT was set for two months after the start date.  Setting the reward per HIT and number of HITs available are interdependent because they determine how much money needs to be set aside for the HIT.  Take the reward per HIT multiplied it by the number of available HITs plus 10% for MTurk and that gives the amount of money that will be taken from the prepaid balance.  **Figure 9** below shows the specific amounts used for the reward per HIT, $0.10, and number of available HITs, 900.  $99 was removed from the prepaid balance before the HITs could be released to the workers.  Once money is transferred to the requester’s prepaid balance, the requester can post a HIT.



$0.100 Reward per Assignment:

x 900

$90.000 Estimated Total Reward:

+ $9.000 Estimated Fees to Mechanical Turk:

**$99.000** Estimated Total Cost:

***Figure 9:*** *Estimated Total Cost Calculation*

The requester has two responsibilities to fulfill between the time a HIT is posted and it expires or is taken down by the requester. The first of these two responsibilities is approving HITs. This process can be done manually or automatically depending on preference. If done manually the requester must check each one of the submissions to see if it was completed adequately before accepting the response. If done automatically, a submission is always accepted after a timeout period. The advantage of the first method is that only high quality data is collected, whereas the advantage of the second method is that a large quantity of responses can be handled over the course of the study. The second responsibility for the requester is to respond to feedback that workers can submit after completing the HIT. This feedback can be positive, but more likely it is negative. In the case of the survey in this research, negative feedback usually pertained to the external Google Forms survey, which could be easily modified at any time. As long as the acceptances and complaints are addressed throughout the HIT campaign, the process of maintaining the job is not difficult to manage.

The two ways a HIT can be terminated are through a manual removal or HIT expiration. The requester has the power to take down the HIT at any time throughout its campaign.  The HIT for this research was removed in this way as there were an adequate number of responses collected before the planned termination date. Otherwise the HIT can be left to expire at its planned date.  Both methods have the same result and the requester is not penalized for removing a HIT early.  Once the HIT has been terminated, the money withheld for that HIT is transferred back to the prepaid balance for the requester. The requester can then contact MTurk support to have the money transferred into an Amazon Payments account if desired.  The HIT is no longer visible to workers at this point and there is no more chance for submissions from that HIT.

## 5 Results

The survey conducted through Mechanical Turk resulted in over 250 total responses. Invalid responses include submissions with incorrect validation numbers, submissions that took the user a shorter amount of time to complete than the total length of the four videos, and submissions that show the user only chose the first response to the required questions. There were 147 valid responses. The responses were then analyzed using a cumulative distribution function. The results of the analysis for each test video can be seen in **Figures 10-13**. The graphs show the users rated QoE for each of the four test videos.

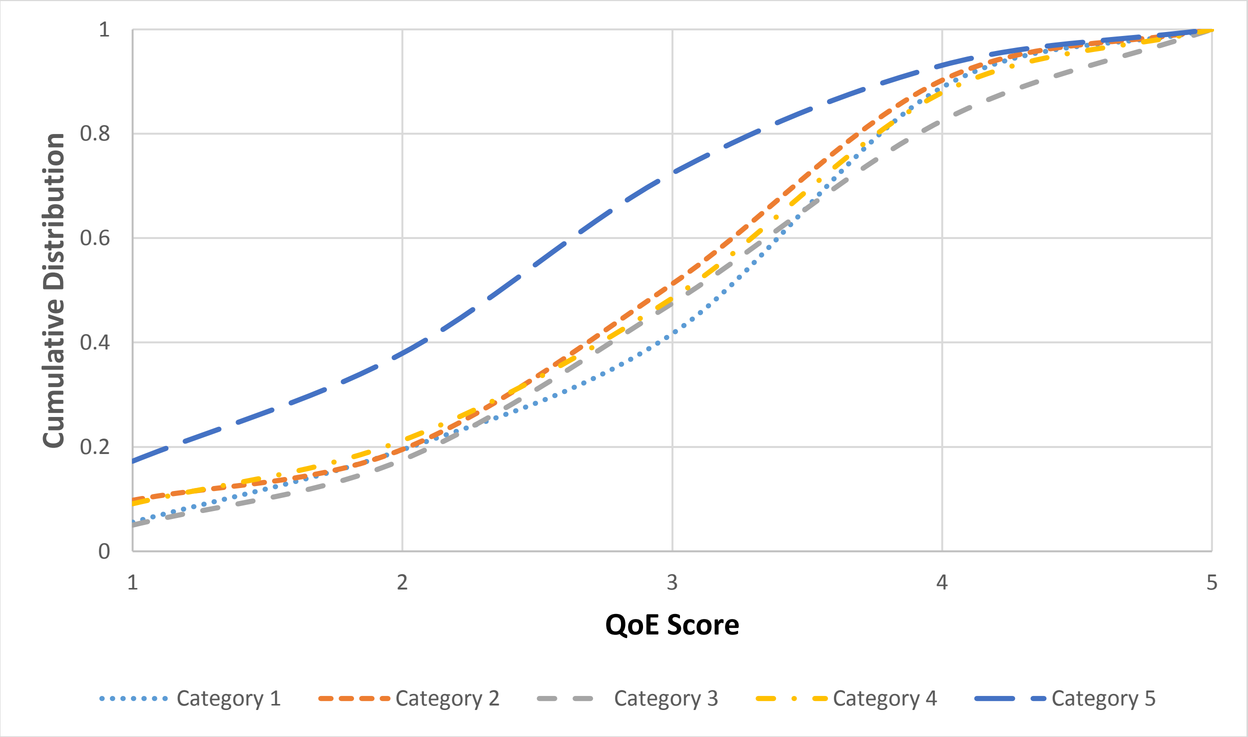


***Figure 10:*** *Netflix Video 1 CDF*

**Figure 10** shows a CDF of the MTurk user's QoE responses for the first Netflix test video. The users viewed one video that is representative of Netflix content and gave it a QoE score afterward. The X-axis represents the user's QoE score with 1 being unsatisfied and 5 being satisfied with his or her experience. The Y-axis is the cumulative distribution of the responses per video category. The users watched one of 5 variations of the same video that fit into the categories described in Section 3.2. Notice that there is little difference among the 5 categories with respect to the cumulative distribution. All the responses follow the same trend independent of the video's category, except for category 5, which has received lower average ratings overall.

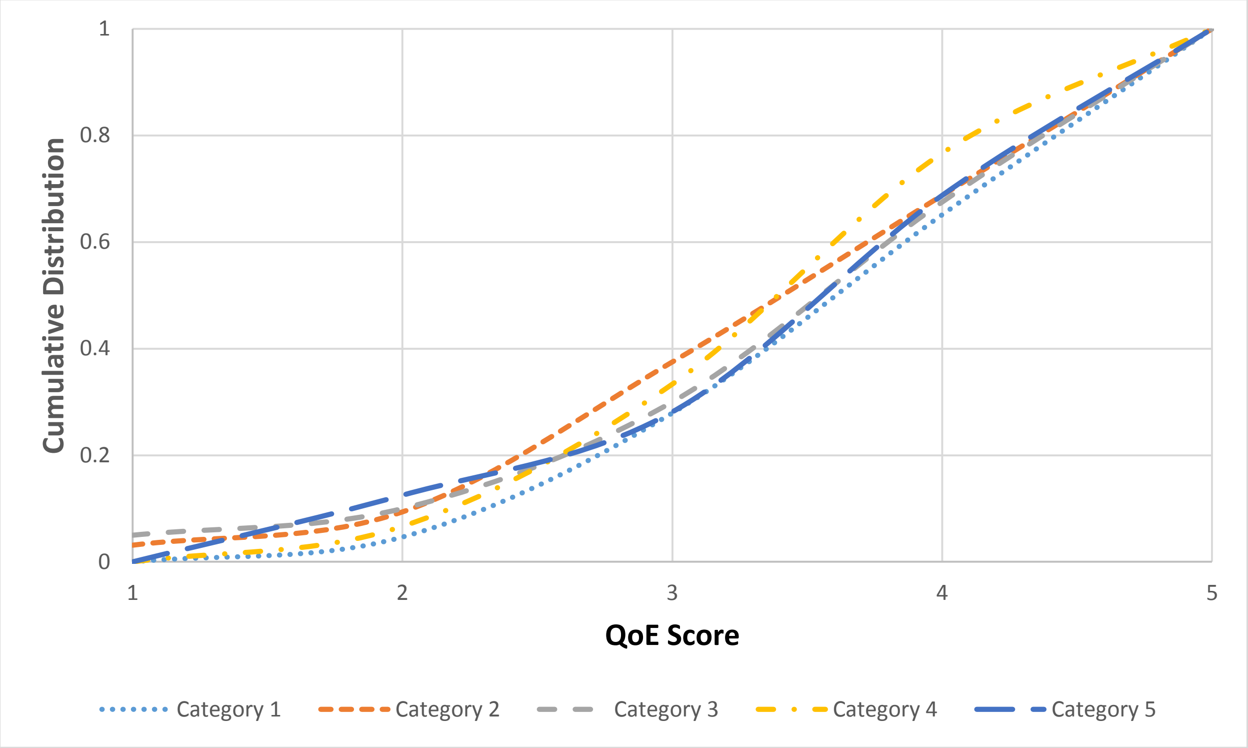


***Figure 11:*** *Netflix Video 2 CDF*

**Figure 11** shows a CDF of the MTurk user's QoE responses for the second Netflix test video. The second video users watched and provided QoE data for is representative of Netflix and has different content from the first. The X and Y-axis are the same as in **Figure 10** and users were given a video from a random category. Notice how the cumulative distribution for each category shows less variance than the first Netflix video. Category 5 does not have a lower average score overall in this video and all categories follow the same trend.

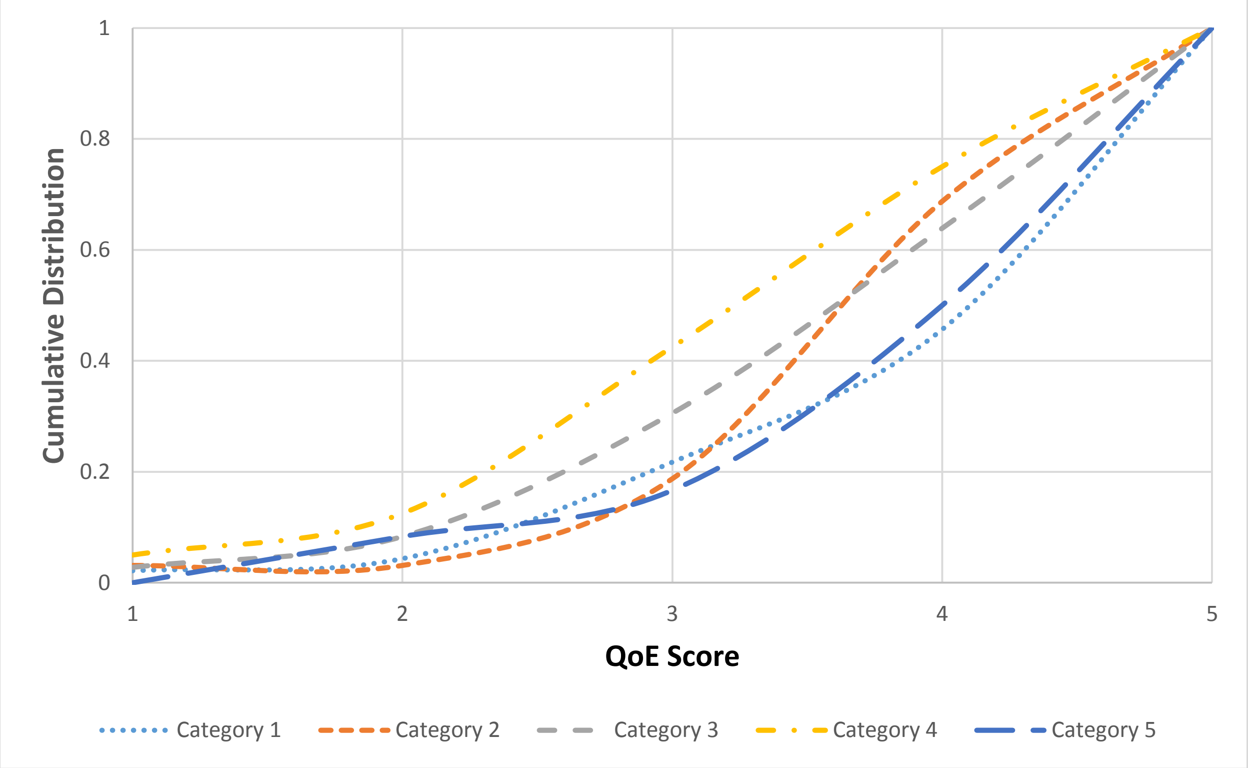


***Figure 12:** YouTube Video 1 CDF*

**Figure 12** shows a CDF of the MTurk user's QoE responses for the first YouTube test video. The axis are the same as the previous two figures, but users now are watching videos that are representative of YouTube content. Notice how there is more variance among the cumulative distributions for each of the categories. Overall, however, the cumulative distributions follow the same trend as before and no clear distinction can be made among the categories.  In order to make a distinction, adjacent categories like 4 and 5 must have similar trends to each other that are different from the remaining categories'.  What is evident though, is that categories 4 and 5, which should receive similar responses because their buffer/ interrupt ratio is similar, are at opposite ends of the cumulative distributions relative to the other categories.

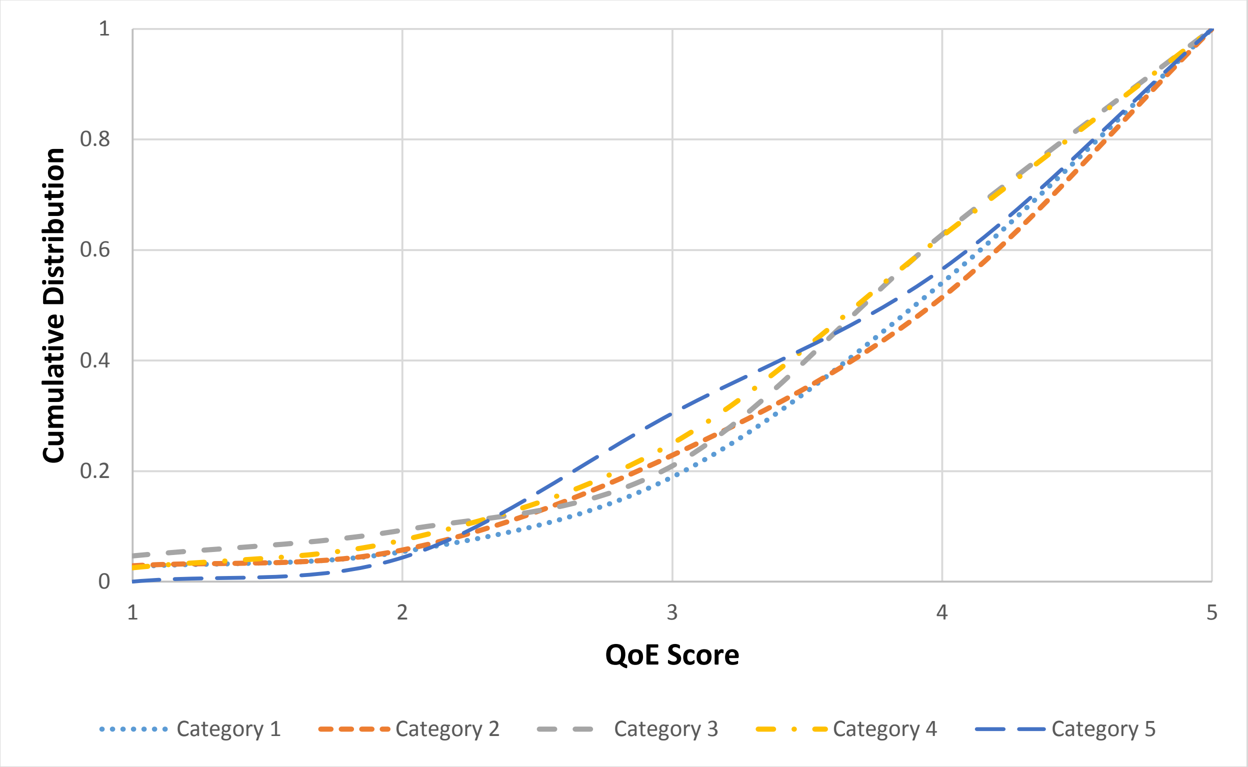


*Figure 13: YouTube Video 2 CDF*

**Figure 13** shows a CDF of the MTurk user's QoE responses for the second YouTube test video. The axis are the same and the data represented is user feedback on the second YouTube video.  Once again there is little variance in user QoE among the 5 categories and the ratings continue to follow the general trend that now can be said to be independent on video content.  There are no outliers in this data for the second YouTube video.

The results of the CDFs for each video shows that regardless of what category a video is in, it receives approximately the same distribution QoE score. This does not support our hypotheses and goes against the intuition that there should be some overall difference in the responses. While there could be many factors that have cause this, the most concerning is the quality of MTurk responses. While invalid MTurk responses have been removed, there is no way to tell for the remaining reposes if the user have answered the questions truthfully and that the MOS responses are representative of their true QoE. Another issue could be that the duration of video playback and video pausing was, even though it was distributed differently, identical in each video. This could imply that the total waiting time, rather than where the waiting time is distributed, affects the quality. Another smaller survey was done on Reddit’s survey page *r/Sample Size*. The survey only received 19 responses; a combination of all four videos’ responses can be seen in **Figure 14** below. There is not a large enough sample size to make any definitive conclusion, however while it is more divergent from the MTurk graphs, it is not divergent enough to suggest there is a significant difference between the different video categories.

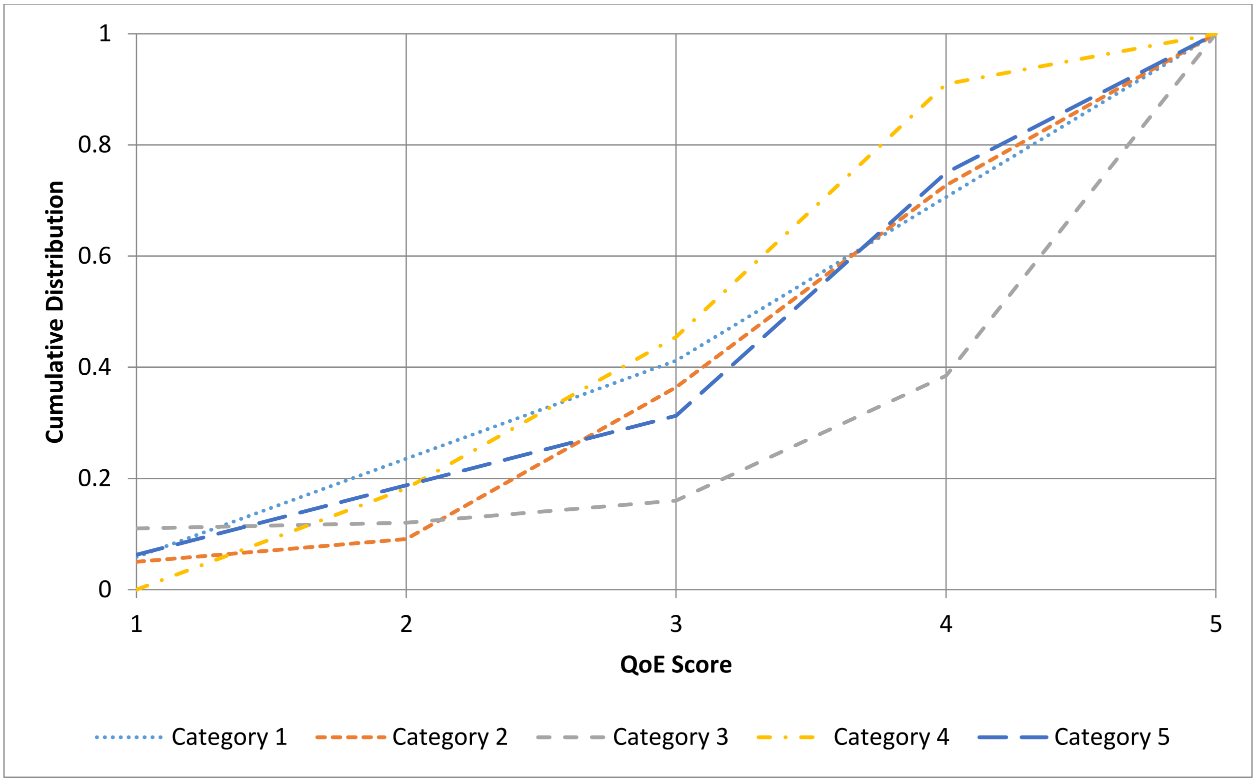


***Figure 14:*** *Cumulative Reddit QoE CDF*

The data collect does not support the hypothesis that there is a difference in user QoE with regards to buffer and interrupt ratio. The data also does not support that users have different expectations based on the content source. This would suggest further study into this topic with other methods than the one used in this study, likely one with in-person subjects.

# 6 Conclusions

This IQP did research to determine if there is a QoE difference, from the user's perspective, between the amount of buffering, and the number of interrupts in a video. To try to determine if there is a difference between the two, and if there is, what that difference is, an online survey was conducted. The survey was built using Google Forms, and was distributed with Amazon's Mechanical Turk. The survey consisted of a series of questions and videos that had a varying number of buffers and interrupts.

After a couple months of the survey on Mechanical Turk, over 250 responses were gathered; of which 147 were deemed to be genuine responses that were used to identify correlations between the buffer/interrupt ratio and QoE ratings. The CDF revealed there was little to no divergence in QoE between the different categories of videos. This means that the hypothesis that states there is a difference in user expectations is not supported. This lack of divergence can be attributed to many factors including flawed methodology, improper incentives for participants, and incorrect hypothesis. Further work needs to be completed to make a definitive claim on the hypothesis.

# 7 Future Work

The analysis of the data from this research has made evident two main areas of improvement. First, the methodology needs to be modified so that it avoids modifying more than one variable at once. Second, a balance must be made between quality and quantity with the Mechanical Turk responses.

## 7.1 Methodology Improvements

Three variables were modified among the four sample videos, initial buffer time, number of interrupts, and content type (YouTube or Netflix).  The results show only a general trend for the responses that cannot be used to gather pertinent conclusions. A potentially more effective study would isolate one variable in order to identify its impact on the data.

The revised methodology could require that users watch a control video with no initial buffer or interrupts and then watch two modified videos for each video provider. The first modified video would have either an initial buffer or interrupts added to it but not both and the second modified video would have whatever modification was not used for the first. The reason for randomizing the order of the videos is to eliminate any bias that may occur due to the order of viewing the videos.  After each video the participant would be asked to rate their QoE.

The advantage of isolating the variables in this way, even though it goes against the tradeoffs that exist in a network, is to pinpoint the effect of the modifications relative to the control video.  If one modification has a bigger impact on the participant’s QoE score than the other then it can be said that the user tolerates the other modification more, which can be directly compared to the hypotheses. The way it is done currently does not allow for the identification of a single variable's impact on the QoE score even if there was a divergence in the data.

## 7.2 Improvements in the Use of Mechanical Turk

Mechanical Turk allows for the customization of worker qualification requirements and reward values so that a requester can have some control over who can accept HITs. By changing these values the requester can choose between the quality and quantity of responses. The responses from the Mechanical Turk users in this study show that some workers do not provide quality responses when given low monetary rewards. The goal of the study was to get a large quantity of results and the quality suffered as a result.

A response was rejected if the participant clearly answered questions dishonestly. These include entering an incorrect number for video identification, answering all QoE questions with the same value, and only selecting the first option on all multiple choice questions. Interestingly the responses from the Reddit survey had no rejected responses despite the fact that these participants took the survey voluntarily with no monetary reward. This implies that there are people who are willing to provide quality responses; it will just take the proper incentives.

In the future, this study could set the worker qualifications and reward values so that quality responses are collected at the expense of reducing the quantity responses. The first step to accomplishing this is to set the worker qualifications so that only "Master Workers" can see the HIT. These are workers who have demonstrated excellence in a particular type of HIT, in this case filling out surveys. These workers earned this qualification by providing quality responses to surveys in the past. These workers generally expect a higher pay for their work so the reward amount needs to be adjusted accordingly. This also serves to limit the number of responses because if the individual cost of an HIT is high, then the total number of HITs is lower given a fixed prepaid balance.

These two proposed solutions together promise quality data that directly correlates with the yet unsupported hypotheses. Consideration needs to be given to the potential lack in quantity of responses, however. This may require a higher budget for paying the more experienced Mechanical Turk workers and it may require a longer campaign time on Mechanical Turk in order to gather responses from an inherently smaller pool of participants. If that is what it takes to gather quality data, then it must be done so that conclusions can be drawn on the hypotheses.

## Works Cited

1. FreeWheel Video Monetization Report Q4 2014. Rep. FreeWheel, 2015. Web. 6 Mar. 2015. <http://www.freewheel.tv/docs/Q4_2014_FreeWheel_Video_Monetization_Report_3.pdf>.

2. U.S. Digital Video Benchmark Adobe Digital Index Q2 2014. Rep. Adobe's CMO.com, 2014. Web. 6 Mar. 2015. <http://www.cmo.com/content/dam/CMO_Other/ADI/Video_Benchmark_Q2_2014/video_benchmark_report-2014.pdf>.

3. J. G. Apostolopoulos, W. Tan, and S. J. Wee. “Video Streaming: Concepts, Algorithms, and Systems.” Technical Report. HP Laboratories Palo Alto, 2002.

4. T. Kim and M. H. Ammar, Receiver buffer requirement for video streaming over TCP. SPIE VCIP Conference 2006 (San Jose, CA, Jan. 2006). <http://www.cc.gatech.edu/~ammar/papers/vcip_final.pdf>.

5. Global Internet Phenomena Report 1H 2014. Rep. Sandvine Incorporated ULC. 2014 Web. Revision: 2014-05-15. Pg. 7 <https://www.sandvine.com/trends/global-internet-phenomena/>.

6. A. Ostaszewska, and S. Żebrowska-Łucyk. "The Method of Increasing the Accuracy of Mean Opinion Score Estimation in Subjective Quality Evaluation." Wearable and Autonomous Biomedical Devices and Systems for Smart Environment Vol. 75, (2010): 315-29. Springer. Web. 6 Mar. 2015. Pg. 328 <http://link.springer.com/chapter/10.1007%2F978-3-642-15687-8_16>.

7. F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, “CrowdMOS: An approach for crowdsourcing mean opinion score studies,” in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, May 2011. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5946971>.

8. J. Xu, L. Xing, A. Perkis, and Y. Jiang, “On the properties of mean opinion scores for quality of experience management,” in In Proc. of Int’l Symp. on Multimedia (ISM), 2011, pp. 500–505. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6123396>.

9. F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. A. Joseph, A. Ganjam, J. Zhan, and H. Zhang. Understanding the impact of video quality on user engagement. In Proc. SIGCOMM, 2011. <https://www.cs.cmu.edu/~xia/resources/Documents/comm254-dobrian.pdf>.

10. A. Finamore, M. Mellia, M. Munafo, R. Torres, and S. G. Rao. Youtube everywhere: Impact of device and infrastructure synergies on user experience. In Proc. IMC, 2011. <http://conferences.sigcomm.org/imc/2011/docs/p345.pdf>.

11. T. De Pessemier, K. De Moor,W. Joseph, L. De Marez, and L. Martens, “Quantifying the influence of rebuffering interruptions on the user’s quality of experience during mobile video watching,” Broadcasting, IEEE Transactions on , vol. 59, no. 1, pp. 47–61, March 2013 <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6323050&tag=1>

12. “Global Internet Phenomena Report 1H 2014.” Rep. Sandvine Incorporated ULC. 2014 Web. Revision: 2014-05-15. Pg. 6 <https://www.sandvine.com/trends/global-internet-phenomena/>.

13. V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z.-L. Zhang, “Unreeling netflix: Understanding and improving multi-cdn movie delivery,” in Proceedings of IEEE INFOCOM, 2012. <https://www.cs.princeton.edu/courses/archive/fall14/cos561/papers/NetFlix12.pdf>.

14. P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: A view from the edge. In IMC, 2007. <http://www.hpl.hp.com/techreports/2007/HPL-2007-119.pdf>

15. Haddad, M., Altman, E., El-Azouzi, R., Jimenez, T., Elayoubi, S. E., Jamaa, S. B., Legout, A., Rao, A., “A Survey on YouTube Streaming Service.” Journal. 2010. Web. <http://www-sop.inria.fr/members/Eitan.Altman/PAPERS/survey_youtube.pdf>.

16. S. Boyer, M. Stron “Best Practices for Improving Survey Participation.” Technical Report. Oracle. 2012. Web. <http://www.oracle.com/us/products/applications/best-practices-improve-survey-1583708.pdf>.