Evaluation of Search Speech - Guidelines

Version 1.0 - December 13, 2017

Instructions

An **eyes-free voice assistant** is an electronic device that can understand voice queries from a user, and give audio responses or take actions on the user's behalf. For example, Google Home is a smart speaker that lets you interact eyes-free with the Google Assistant. An eyes-free device does not have a screen, but may be connected to other devices in the user's home. For example, an eyes-free device may be able to connect to the TV to play a video. For this task, the term **eyes-free** is used for devices that have no screens or keyboards (or very limited screens and keyboards), so the primary method of interacting with the device is talk to it, and listen to audio responses.

In this task, you will see a query a user might speak to the Google Assistant, and a corresponding response from the Assistant. It is very important for you to represent the users in the locale you evaluate. The response might include audio spoken by the device, a text description of an action taken by the device, or both, or sometimes a video played on TV. Your job is to provide feedback about the response from the Assistant. Specifically, you will give feedback about the usefulness of the response, and for audio responses, the quality of the speech. More detail about each rating is given below.

Needs Met Rating

Please refer to the General Guidelines for instructions on how to rate Action and Answer results using the Needs Met scale. Below are some additional examples of responses from an eyes-free device.
Click to show/hide examples for Answers

Query	Spoken Response	Rating	Discussion
[how tall was charles darwin?]	Charles Darwin stood about 5 feet, 11 1/2 inches tall.	Fully Meets	This spoken response fully answers the user's query.
[william blake]	According to example.com, William Blake was an English poet, painter, and printmaker.	Highly Meets	This spoken response would meet the information need stated by the query. Some users might want additional information, and that is made available on the referenced website. The user will receive a link to the

			specific page.
[what will the weather be like this weekend?]	It will be 69 degrees and cloudy	Moderately Meets to Slightly Meets	The response contains some useful information, but the query seeks a weather forecast for the entire weekend. Most users would want a more detailed response.
[who is the president of the united states?]	According to example.com, the president of the united states is the elected head of state of the united states.	Slightly Meets	Most users issuing this query want to know the name of the current president, not the definition of the office. This response would be helpful for few users.
[will it rain this evening?]	I'm not sure how to help with that.	Fails to Meet	The device failed to answer the query. No users would be satisfied with this response.

Click to show/hide examples for Actions

Query	Response	Rating	Discussion
[play beethoven ninth symphony on spotify]	Action Response: Play Media Artist: Beethoven Song: Ninth Symphony Content URL: URL	Fully Meets	The device plays the requested song from the correct source. This Fully Meets the user need.

[play the magic flute]	Action Response: Play Media Artist: Mozart Music: The magic flute Content URL: link to an amateur representation of The Magic Flute.	Highly Meets	The device played the opera by Mozart as performed by an amateur orchestra, taken from YouTube. Some users might want a professional version.
[play jazz]	Action Response: Play Media Artist: some popular jazz artist.	Moderately Meets to Slightly Meets	The device plays a certain jazz song, which would be satisfying to some users. Most users would prefer a jazz playlist to be played.
[play music video on my tv]	Action Response: Play Video on TV from YouTube that only contains the song, and the lyrics.	Slightly Meets	An audio-only version of a song was played on TV. Although some users could be satisfied by this audio-only version, most users would prefer an official music video.
[play mumford and sons reminder]	Action Response: Set a Reminder Time: Please specify a time	Fails to Meet	The user wanted to play a specific song, and the device instead set a reminder. No users would be satisfied with this response.

Speech Quality Rating

For audio responses, you also need to rate the speech quality of the spoken result. For this speech quality rating, do not consider the factual correctness of the response, but instead consider each of the following factors below to determine if the speech quality is acceptable.
Every task will ask you about the following two factors:

- **Length:** Was the response of an appropriate length matching the complexity of its content? Would it have been beneficial for the user if it were either more concise or more detailed?
- **Formulation:** Was it grammatical? Was the response formulated in a way you would expect a native speaker to formulate it? As opposed to a machine or someone not fully fluent in the language. When the answer included an attribution, was the source of the content clear and understandable?

Additionally, some tasks will ask you about one additional factor:

- **Elocution:** How accurate were the pronunciations of each word spoken in the response? Was the intonation of the voice natural throughout the response? Was the speed at which the response was spoken appropriate? This could be due to an awkward rhythm in the spoken response, or words slightly mispronounced.

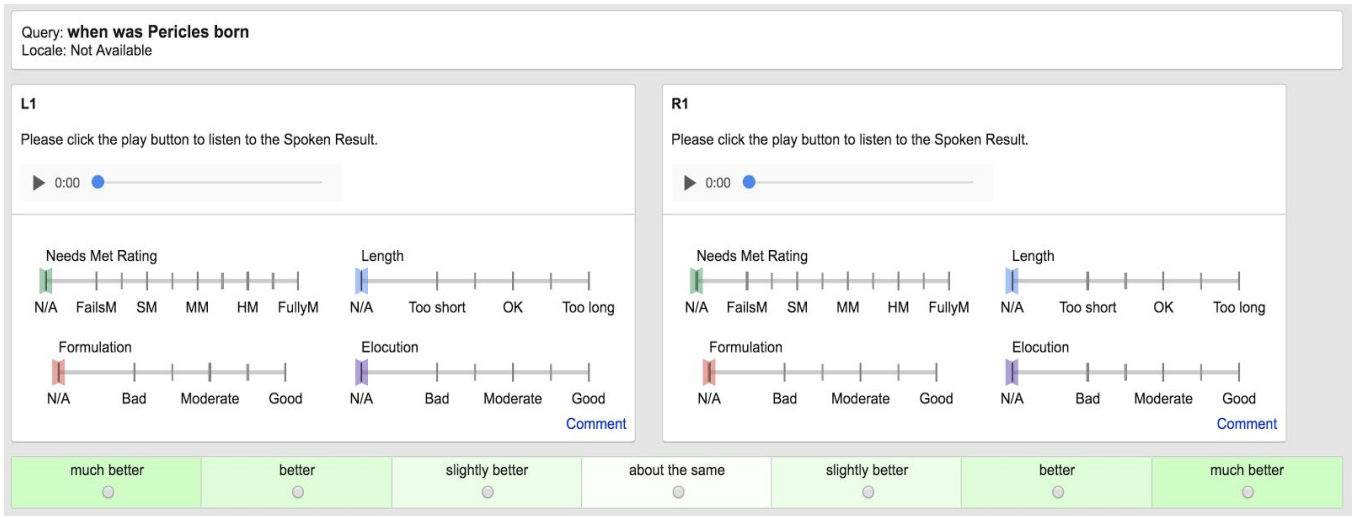
Click to show/hide Speech Quality rating examples

Query	Spoken Response	Rating	Discussion
[beethoven pathetique sonata]	Playing Beethoven's pathetique sonata	Length=**OK** Formulation=**Good** Elocution=**Good**	The response is well-formed with no issues.
[how far is alpha centauri from the sun]	Alpha Centauri is 4.367 light years from earth.	Length=**OK** Formulation=**Good** Elocution=**Good**	The response is well-formed with no issues.
[where did einstein go to college]	On the website example.com, they say: Joseph Einstein is expected to attend Harvard in Fall 2017.	Length=**OK** Formulation=**Good** Elocution=**Good**	The response is incorrect, since the question is likely asking about Albert Einstein, but the spoken response is concise and has no obvious errors. All three factors are ok.
[is a pregnancy test accurate]	On the website example.com, they say: However, recent research indicates that if a woman has missed a period, then many home pregnancy tests are not sensitive enough and cannot diagnose pregnancy.	Length=**A bit long** Formulation=**Moderate** Elocution=**Good**	The formulation of the response has a minor problem, in that it has an erroneous leading phrase "however". Length is a bit long as the response could have been more concise.

[what's the highest interest rate on a car]	On the website example.com, they say: Independent Consultants reported that if buyers have a credit score below 550, the interest rates on a new vehicle loan can be as low as 12% and on a used vehicle loan they can be as low as 17%, according to McGriffiths as reported.	Length=**Too Long** Formulation=**Bad** Elocution=**Moderate**	Inside the snippet there is both a leading attribution and an attribution at the end. This makes the passage poorly formulated. McGriffiths is out of context. The answer is also clearly too long.
[what is pasteur best known for]	Here's a summary from the website example.com: Louis Pasteur found microbes sour alcohol proposed pasteurization, kills bacteria with heat.	Length=**OK** Formulation=**Bad** Elocution=**Good**	While the answer is good and has an appropriate length, the text reads somewhat ungrammatical.
[what does BMI stand for]	Body Mass Index. According to Wikipedia: BMI stands for Body Mass Index and it's becoming a universal tool to measure body "fatness" even though it doesn't actually measure body fat like using a caliper or underwater weighing.	Length=**Too Long** Formulation=**Good** Elocution=**Good**	This response is too long, and provides additional, irrelevant information.
[what is the outer layer of your skin called]	According to the website example.com: Called the epidermis, the outer layer, and called the dermis, the inner layer.	Length=**OK** Formulation=**Bad** Elocution=**Good**	The response is not a complete sentence. This should be rated low on formulation.

[how old is Achilles during Iliad]	According to the website example.com: In that movie about Troy Achilles is depicted as being 27 28th years old as Agamemnon is 26 27th years old, Hector is about 30 31th years old and Tiresias is 54 55th years old.	Length=**Too Long** Formulation=**Bad** Elocution=**Bad**	27/28 years old gets read as "27 twenty eighth years old" (same for the other numbers). The ages of all the characters make it harder to follow and definitely longer than it should be. Elocution for the numbers is also poor.
[how old is Achilles during Iliad]	On the website example.com, they say: Achils sounds old on 20, she's 17 or 18.	Length=**OK** Formulation=**Bad** Elocution=**Bad**	This response has more than one significant problem. The name Achilles is mispronounced, the phrase "Achils sounds old on 20" is confusing.
[who is the tallest man in the world]	Here's a summary from the website example.com: Robert Pershing Wadlow	Length=**Too Short** Formulation=**Good** Elocution=**Good**	The user probably wants to know not just who the tallest man is, but the height. The response is too short.
[world population]	The population of the world is 6,999,999,989.	Length=**OK** Formulation=**Bad** Elocution=**Bad**	The number in the response would be tedious when spoken aloud. It should be rounded.

Note: You may also see a side-by-side rating question. In this case, please rate which side better meets the user need.



found on 192.168.1.1 and 192.168.0.1