JSLHR

Research Article

A Potential Bias in Subjective Ratings of Mental Effort

Travis M. Moore[a] and Erin M. Picou[a,b]

Purpose: Subjective reports of listening effort are frequently inconsistent with behavioral and physiological findings. A potential explanation is that participants unwittingly substitute an easier question when faced with a judgment that requires computationally expensive analysis (i.e., heuristic response strategies). The purpose of this study was to investigate whether participants substitute the question "How did I perform?" when asked "How much effort did that take?".

Method: Participants completed 2 sets of online surveys containing a text-based, multiple-choice synonym task. Expected performance and mental effort were manipulated across sets in 4 experiments, using a visual masking technique shown to correlate with speech-reception-testing in noise. Experiment 1 was designed to yield stable accuracy and differing effort across sets. Experiment 2 elicited differing accuracy and stable effort. Experiments 3 and 4 manipulated accuracy and performance in opposite directions. Participants included 273 adults (aged 19–68 years, *M* = 38.4 years).

Results: Experiment 1 revealed no influence of perceived performance on ratings of effort when accuracy was stable. Experiment 2 showed that ratings of effort differed inversely with ratings of performance (lower performance and increased effort). Experiments 3 and 4 also demonstrated that participants rated effort in a manner inversely related to performance, regardless of the effort inherent in the condition.

Conclusions: Participants likely substitute an easier question when asked to rate the multidimensional construct of mental effort. The results presented here suggest that perceived performance can serve as a ready heuristic and may explain the dissociation between subjective measures of listening effort and behavioral and physiological measures.

Peripheral hearing status is only one of many factors that influence the subjective experience of hearing loss, which in turn affects a variety of issues such as perceived hearing handicap and hearing aid acceptance (Gatehouse & Noble, 2004; Pichora-Fuller & Singh, 2006). Understanding the experience of hearing loss must also take into account higher cognitive functions, including the interaction between sensory perception and cognitive processing (Gatehouse & Noble, 2004). One such phenomenon that has received significant attention in the literature is listening effort (Gagné, Besser, & Lemke, 2017; Pichora-Fuller et al., 2016). Listening effort typically refers to the cognitive resources required for understanding speech, with the implication that challenging listening conditions increase the necessary resources for successful speech processing (Fraser, Gagné, Alepins, & Dubois, 2010; McCoy et al., 2005). More specifically, listening effort appears to be a multidimensional phenomenon, dependent on factors such as audibility (Hornsby, 2013), listening environment (Picou, Moore, & Ricketts, 2017), psychophysiological state (Picou & Ricketts, 2014), and other cognitive abilities (Zekveld, Kramer, & Festen, 2011). The interaction of these factors determines the amount of effort, and thereby cognitive resources, allocated to a listening task.

Attempts to quantify the complex phenomenon of listening effort have understandably produced a wide array of laboratory tests, including physiological, behavioral, and subjective measures. Physiological tests use biological data to establish an index of listening effort. Kahneman (1973) suggested that, because physiological arousal covaries with mental effort, physiological techniques can serve as measures of effort. Currently, pupillometry has become an accepted physiological measure of listening effort (Engelhardt, Ferreira, & Patsenko, 2009; Zekveld, Kramer, & Festen, 2010), joined by an array of additional (electro)physiological measures, such as skin conductance, heart rate variability, electromyography, and electroencephalography (e.g., Mackersie & Cones, 2011; Mackersie, MacPhee, & Heldt, 2015; Obleser & Kotz, 2011).

[a]Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN
[b]Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN

Correspondence to Travis M. Moore: travis.m.moore@vanderbilt.edu

Editor-in-Chief: Sean Redmond
Editor: Daniel Fogerty

Received December 5, 2017
Revision received March 28, 2018
Accepted May 21, 2018
https://doi.org/10.1044/2018_JSLHR-H-17-0451

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

One behavioral method of quantifying listening effort is the dual-task paradigm (Broadbent, 1958; Rabbitt, 1966), wherein participants simultaneously perform two separate tasks, with instructions to maintain performance on the primary task (e.g., speech recognition). The conventional interpretation is that increasing the mental demand of the primary task eventually exceeds the available resource capacity of an individual, resulting in withdrawal of resources from the secondary task. The resultant change in secondary task performance is considered a behavioral index of listening effort (e.g., Fraser et al., 2010; Gagné et al., 2017; Gosselin & Gagné, 2011b; McCoy et al., 2005).

Subjective measures of listening effort typically include both validated and unvalidated questionnaires and rating scales (Fraser et al., 2010; Hart & Staveland, 1988; McNair, Lorr, & Droppleman, 1971; Picou et al., 2017; Zijlstra, 1993). Validated questionnaires have the benefit of documented psychometric properties and can provide normative values, whereas custom questions are not typically subjected to formal checks. However, unvalidated questions can serve as focused assessments of a particular attribute under conditions specific to a study (see Moore, Key, Thelen, & Hornsby, 2017, for further discussion). While assessing listening effort can involve sophisticated yet indirect behavioral and physiological measures, the most direct insight into the mental state of a human listener might simply be to rely on participants' subjective report. Subjective questionnaires are noninvasive, low-risk measures that can be shared across laboratories without the need for specialized software or equipment. As the least expensive and most easily administered test, subjective questionnaires are frequently included de facto in studies of listening effort. In addition, subjective listening effort has been the only measure of listening effort reported in some studies (e.g., Brons, Houben, & Dreschler, 2013; Luts et al., 2010; Rennies, Schepker, Holube, & Kollmeier, 2014). The interpretation of such investigations relies on a comprehensive understanding of the factors that affect subjective ratings and also an understanding of the relationship between subjective reports and objective or behavioral measures of effort.

Disagreement Across Methodologies

There is general agreement between behavioral and physiological measures, which suggests that these techniques are sensitive to similar aspects of listening effort (e.g., Mackersie et al., 2015; Zekveld et al., 2010; but see Hicks & Tharpe, 2002). Conversely, it has been shown that subjective measures of listening effort commonly agree with neither behavioral nor physiological measures (e.g., Desjardins & Doherty, 2013; Gosselin & Gagné, 2011b; Hicks & Tharpe, 2002; Hornsby, 2013; Larsby, Hällgren, Lyxell, & Arlinger, 2005; Mackersie & Cones, 2011; Mackersie et al., 2015; Miyake, 2001; Picou et al., 2017; Wilson & Sasse, 2001; Yeh & Wickens, 1988; Zekveld et al., 2010, 2011). For instance, Zekveld and Kramer (2014) asked participants to indicate their effort during speech recognition tasks with various levels of difficulty (masking). Using pupillometry, the authors noted reports of high effort and overall smaller pupil dilation in the most difficult conditions. These results demonstrate physiologic evidence of disengagement and low effort but subjective reports of high effort in the same conditions. The lack of agreement between subjective measures and behavioral or physiological measures of listening effort undermines current understanding of the phenomenon, yet the reason for the disagreement remains uncertain.

One interpretation of uncorrelated findings between subjective measures and other methods is that subjective report is sensitive to different parameters of listening effort compared with behavioral and physiological measures. The natural question that follows is "To what then are subjective measures sensitive?" This question may initially seem more daunting than warranted, in part, because of its basis on the apparent validity of simply asking human volunteers to report on their experiences. Johanssen et al. (1979, p. 105) have been quoted to summarize this rationale as follows: "If the person <u>feels</u> loaded and effortful, he <u>is</u> loaded and effortful whatever the behavioral and performance measures may show" (original emphasis). However, perhaps the relationship between asking a participant a question and receiving a scientifically valid answer is not as straightforward as it seems.

Heuristics and Biases

In their seminal work, Tversky and Kahneman (1974) offered a theory that sought to explain how people assign value to uncertain quantities, such as happiness. They suggested that people do not base their judgments on strictly logical assessments but rather rely on a small set of strategies that reduce the effort and complexity of decision making. They referred to such strategies as judgment heuristics. More specifically, heuristics have been defined as simple strategies that replace complex processes to produce imperfect but acceptable solutions that require minimal cognitive effort (Newell & Simon, 1972; Simon, 1990). Shah and Oppenheimer (2008) suggest that heuristic response strategies minimize effort by examining less information (e.g., fewer cues and alternatives) and by reducing the load on memory storage and retrieval. Consistent with this view and most pertinent to the current study, Kahneman and Frederick (2002) described a general heuristic phenomenon they termed *attribute substitution*: When people are faced with a difficult question, they often, unwittingly, substitute an easier question that they answer instead. For instance, Kahneman and Frederick give the example of a professor who has just watched a candidate's job talk. In deciding whether or not to hire the applicant, the professor is asked to answer the question, "How likely is it that this candidate could be tenured in our department?", but instead comes to a decision by answering the easier question, "How impressive was the talk?". That is, rather than consider the candidate as a whole, the hiring professor made a decision based on the more easily quantifiable and easily accessible impression of the job talk.

In the illustration above, the target attributes to be judged in hiring a new faculty member are clear (e.g., number of first-author publications, strong history of independent funding), but the case of judging listening effort becomes even more difficult because there are no readily available criteria for judging. The lack of a familiar scale for estimating listening effort renders a seemingly simple question much more difficult to answer, creating a situation well suited to the use of automatic, heuristic judgments.

A Potential Heuristic Attribute for Listening Effort

We suggest that subjective ratings of listening effort are susceptible to the same processes described above, which lead participants to substitute the target question regarding listening effort with an easier question. Doing so biases participant responses toward a related yet more easily judged domain. Finding a common heuristic attribute used by participants when rating listening effort could inform study design and subjective questionnaire formation to avoid or account for the bias. Such measures might reconcile the discrepancy between subjective report and other techniques, solidifying the current understanding of listening effort and perhaps eventually aiding in the establishment of a universal definition.

The current challenge is to identify and test potential biases. This study hypothesizes that the question "How well did I perform?" is substituted for the question "How much effort did that take?". This hypothesis is consistent with the findings from Picou et al. (2017), who found that two subjective questions related to listening effort were actually better correlated with task performance (word recognition) than behavioral measures of listening effort (see also Fraser et al., 2010; Larsby et al., 2005; Picou & Ricketts, 2018; Zijlstra, 1993). This hypothesis also satisfies the three conditions provided by Kahneman and Frederick (2002) for a scenario likely to give rise to attribute substitution. First, the target attribute must be relatively inaccessible. As described above, listening effort fulfills this condition as a multidimensional, psychophysiological phenomenon with no established rating scale. Second, a semantically and associatively related attribute must be highly accessible. Unlike ratings of listening effort, subjective and behavioral ratings of speech recognition are often highly correlated (e.g., Cienkowski & Speaks, 2000; Cox, Alexander, & Rivera, 1991; Larsby & Arlinger, 1994), suggesting that performance is a readily accessible quantity. Third, the substituted attribute must be reasonable. In a manner similar to the hiring committee judging the likelihood of tenure based on the easy question of the quality of the job talk, we suggest that perceived performance on a task might bias ratings of perceived mental effort.

Purpose

The purpose of this study was to investigate whether participants substitute an easier question ("How did I perform?") when faced with the more difficult question of "How much effort did that take?". To accomplish this, we used scenarios where a participant's task accuracy either would or would not be expected to bias subjective ratings of effort. If subjective ratings can be biased by perceived task accuracy, investigations into mental effort should take this bias into consideration in both study design and data interpretation. The finding that subjective ratings can be biased by performance would also help clarify the noted dissociation in the literature between subjective ratings of effort and behavioral or objective indices of effort. By evaluating a variety of experimental scenarios, the results of this study will advance our understanding of which experimental conditions, if any, make use of perceived performance to bias subjective ratings of mental effort.

To reach a large number of individuals, we made use of crowdsourcing software to post online surveys. To avoid the uncontrollable variables of delivering audio over the Internet to personal computers, we based this study on a visual analog of the speech recognition threshold (SRT) test, called the Text Reception Threshold Test (TRT; Kramer, Zekveld, & Houtgast, 2009; Zekveld, George, Kramer, Goverts, & Houtgast, 2007). As discussed in Zekveld et al. (2007), the TRT is sensitive to the "modality-aspecific" cognitive functions that are common between understanding verbal language in background noise and understanding written language embedded in visual noise. They showed that the TRT was significantly associated with the SRT ($r = .54$, $p < .01$). Thus, although the TRT stimuli in the current experiment used the visual modality, the task was also sensitive to top–down processes governing linguistic ability across the auditory and visual domains. Although the immediate impact on traditional listening effort will require further investigation, the findings presented here are readily applicable to the field because of the significant overlap between factors that contribute to speech recognition across vision and audition (e.g., Humes, Burk, Coughlin, Busey, & Strauser, 2007; Zekveld & Kramer, 2014; Zekveld, Pronk, Danielsson, & Rönnberg, 2018).

To test the hypothesis that performance can serve as a heuristic substitution for subjective ratings of effort, separate cohorts of participants were asked to rate either performance, effort, or work after completing one of three online experiments using visual stimuli with various levels of masking. Each online experiment consisted of two multiple-choice question sets, which were systematically varied to result in specific levels of performance and effort. Table 1 summarizes the experimental designs and expected outcomes. The experiments were designed, respectively, to result in question sets that produced the same performance but different amounts of effort (Experiment 1), different levels of performance but the same amount of effort (Experiment 2), and different, and opposite, levels of performance and effort (Experiments 3 and 4). The third and fourth experiments were conceptually identical, except that different masking levels were used to confirm the validity of experimental manipulations. Expected outcomes of the experiments were that participants would (a) rate mental effort when performance could not serve as a heuristic, (b) use change in performance as a heuristic substitution even when expected effort did not change, and (c) use change in performance as a heuristic attribute when both performance and expected effort changed. An

Table 1. Design of each experiment.

Experiment 1: expect effort lower in Set 1 than Set 2; expect accuracy to be similar in Sets 1 and 2							
Set 1				Set 2			
Number of items	Stimulus visibility	Expected effort	Expected accuracy	Number of items	Stimulus visibility	Expected effort	Expected accuracy
8	70%	Low	100%	8	46%	Moderate	100%
2	28%	Skip	25%	2	28%	Skip	25%
Total		*Low*	*85%*	*Total*		*Moderate*	*85%*

Experiment 2: expect accuracy lower in Set 1 than Set 2; expect effort to be similar in Sets 1 and 2							
Set 1				Set 2			
Number of items	Stimulus visibility	Expected effort	Expected accuracy	Number of items	Stimulus visibility	Expected effort	Expected accuracy
4	70%	Low	100%	2	70%	Low	100%
4	46%	Moderate	100%	2	46%	Moderate	100%
2	28%	Skip	25%	6	28%	Skip	25%
Total		*Low–moderate*	*85%*	*Total*		*Low–moderate*	*55%*

Experiment 3: expect accuracy lower in Set 1 than Set 2; expect effort to be lower in Set 1 than Set 2							
Set 1				Set 2			
Number of items	Stimulus visibility	Expected effort	Expected accuracy	Number of items	Stimulus visibility	Expected effort	Expected accuracy
4	70%	Low	100%	8	46%	Moderate	100%
6	28%	Skip	25%	2	28%	Skip	25%
Total		*Low*	*55%*	*Total*		*Moderate*	*85%*

Experiment 4: expect accuracy lower in Set 1 than Set 2; expect effort to be lower in Set 1 than Set 2							
Set 1				Set 2			
Number of items	Stimulus visibility	Expected effort	Expected accuracy	Number of items	Stimulus visibility	Expected effort	Expected accuracy
4	70%	Low	100%	8	46%	Moderate	100%
6	7%	Skip	25%	2	7%	Skip	25%
Total		*Low*	*55%*	*Total*		*Moderate*	*85%*

Note. Each set included 10 test items, which varied based on the amount of masking covering each stimulus. The table displays the number of items in each set that has a particular visibility. On the basis of the visibility, the expected effort and expected accuracy are displayed. Expected effort and accuracy are displayed across the total set of 10 items (displayed in italic font).

additional question using the word "work" was included as a similar alternative to the "effort" question, with instructions that overtly mentioned that work is different from accuracy. This was done to test the hypothesis that more explicit instructions might promote more effortful rating strategies and reduce the use of heuristic substitutions.

General Method

Participants

Participants were recruited via Amazon's Mechanical Turk (Mturk). Mturk is a crowdsourcing website where anonymous online workers complete tasks on the Internet. Mturk has been shown to be a reliable tool for conducting research in psychology (Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013) and also audiology (Barber & Lee, 2015; Singh, Lau, & Pichora-Fuller, 2015). Using Mturk profile controls, the surveys were only available to workers at least 18 years old and living in the United States. If these criteria were met, participants then answered further eligibility questions before beginning the survey. Specifically, participants had to be native English speakers, with normal or corrected visual acuity and color vision. All participants reported having at least a high school diploma, and 89.7% of the participants reported having completed some level of college. A power analysis revealed that 14 participants per cohort were required to detect small-to-medium effects with 80% power, significant at an α of .05. Participant demographics for each experiment are displayed in Table 2. Table 3 displays the number of surveys

Table 2. Participant demographics across all three experiments.

Experiment	Total *N*	Age, *M* (*SD*)	Age range	% Female
Experiment 1				
Performance	16	42.4 (16.0)	22–66	62.5
Effort	16	39.2 (17.4)	20–65	62.5
Work	14	33.0 (8.9)	22–50	57.1
Experiment 2				
Performance	23	33.4 (10.4)	21–55	52.2
Effort	24	33.6 (10.2)	19–59	58.3
Work	24	37.4 (13.1)	21–68	50.0
Experiment 3				
Performance	28	37.6 (12.7)	19–68	46.4
Effort	24	34.9 (11.3)	19–64	54.2
Work	18	38.4 (14.9)	22–63	38.9
Experiment 4				
Performance	41	40.5 (11.4)	20–62	48.8
Effort	45	37.9 (11.2)	21–63	66.7

Note. Participants answered subjective ratings of either performance, effort, or work in each experiment.

completed, the number excluded for failing the catch trials, and the number of surveys excluded due to suspicious data (performing lower than 50% correct in either set). All testing was completed with approval from Vanderbilt University Medical Center's Institutional Review Board. Participants were compensated for survey completion.

Task and Stimuli

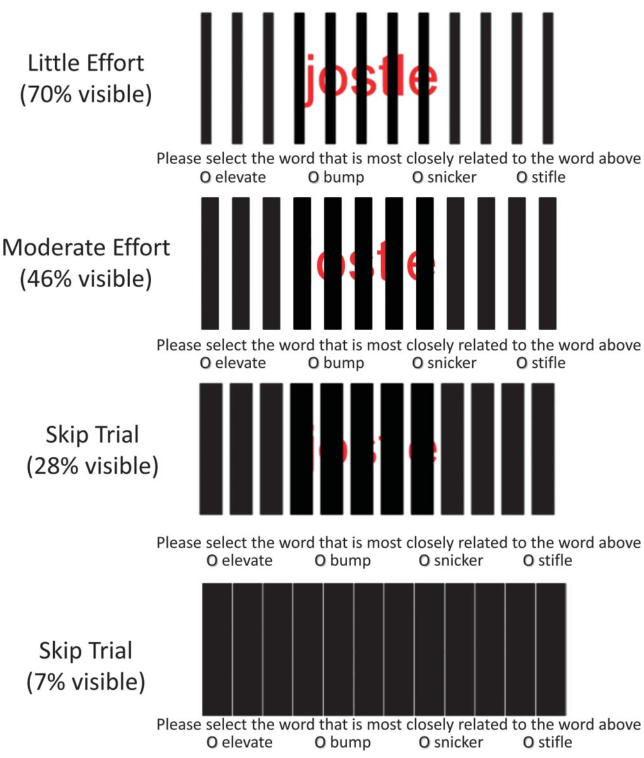
The task consisted of an online survey containing four sections: (a) demographic information, (b) catch trials, (c) multiple-choice questions, and (d) subjective ratings. First, the demographic questions included age, highest level of education, gender, and state of residence. Second, catch trials were designed to identify automated workers or workers who did not follow instructions. Catch trials consisted of decoy instructions followed by target instructions later in the question, requiring careful reading of the prompt. Data from participants who did not correctly answer catch trials were excluded from further analysis.

Third, participants completed the main task, which consisted of multiple-choice, synonym-matching questions. Each trial consisted of a partially masked target word (TRT stimulus), brief instructions, and four unmasked, multiple-choice response options (see Figure 1). Due to the short length typical of Mturk surveys, we limited the task to two sets of 10 questions each (20 questions in total). To manipulate percent correct across question sets, trials consisted of TRT stimuli that were either partially masked or masked to the extent that the word could not be read. It was therefore important that partially masked stimuli produced near-ceiling performance, so that the number of illegible stimuli dictated the percent correct. To achieve near-ceiling performance, all target words were chosen from the Comprehensive Assessment of Spoken Language (Carrow-Woolfolk, 1999), which were normed to be recognizable by individuals 7–21 years old, and were balanced for difficulty across surveys. (Test items are listed in the Appendix.) In this way, if the target word could be read, the trial was likely to be answered correctly. Although this design permitted control over task accuracy, the task was also designed to elicit noticeable mental effort. To this end, the amount of visual masking was varied, and participants were asked to match the target word with a synonym.

Table 3. Number of surveys completed, discarded for failing catch trials, and discarded for suspicious performance (answering fewer than 50% correct in the easiest condition) for each experiment.

Experiment	Surveys completed	Failed catch trials	Suspicious performance	Used for analysis
Experiment 1				
Performance	28	7	5	16
Effort	28	7	5	16
Work	20	0	6	14
Experiment 2				
Performance	33	6	4	23
Effort	30	5	1	24
Work	31	2	5	24
Experiment 3				
Performance	35	5	2	28
Effort	32	5	3	24
Work	32	7	7	18
Experiment 4				
Performance	62	7	10	45
Effort	44	3	0	41

Figure 1. Example stimuli that reflect words that were mostly visible (top), moderately masked (top middle), nearly impossible (bottom middle), and impossible (bottom).


Little Effort
(70% visible)
jostle
Please select the word that is most closely related to the word above
O elevate O bump O snicker O stifle
Moderate Effort
(46% visible)
Please select the word that is most closely related to the word above
O elevate O bump O snicker O stifle
Skip Trial
(28% visible)
Please select the word that is most closely related to the word above
O elevate O bump O snicker O stifle
Skip Trial
(7% visible)
Please select the word that is most closely related to the word above
O elevate O bump O snicker O stifle

All target words (TRT stimuli) were created using custom code in *R* Version 3.3.2 (R Core Team, 2016). The TRT word was printed in red and masked by equal-width, equally-spaced, black bars, based on work by Zekveld et al. (2007). Words were visually obscured with one of four levels of masking (i.e., bar width): 70% visible (see Figure 1, top panel), 46% visible (see Figure 1, second panel), 28% visible (see Figure 1, third panel), or 7% visible (see Figure 1, bottom panel). Threshold procedures and practice lists were not used in this study; masking values were chosen based on values reported as being well below threshold (7% and 28% visible), near threshold (46% visible), and well above threshold (70% visible) in Zekveld et al. (2007). It was expected that no mental effort related to deciphering the TRT word or choosing a synonym would be exerted with 7% or 28% visibility because participants would simply skip over these clearly impossible trials ("skip trials"). Such behavior is consistent with previous findings of reduced effort in difficult or impossible listening situations (e.g., Wu, Stangl, Zhang, Perkins, & Eilers, 2016; Zekveld & Kramer, 2014). That participants randomly guessed at skip trials was later confirmed by the pattern of responses (see Discussion).

The multiple-choice, synonym-matching task used the TRT stimuli described above as target words, to be matched with one of the four unmasked response options (see Figure 1). Participants were instructed to "choose the word that is most related to the word pictured above" and to "take your best guess for all questions." The four multiple-choice options (each a single word) were printed in a standard, black typeface and appeared next to option buttons directly below the target TRT word. Response options were those paired with the test words in the published Comprehensive Assessment of Spoken Language subtest.

The parameters of task accuracy and required effort were strategically varied between the two sets of 10 questions. Percent correct was controlled by the number of skip trials (28% visible), whereas required effort was determined by the ratio of little- to moderate-effort stimuli (70% and 46% visible, respectively). There were four such manipulations (see Table 1). Experiment 1 held expected task accuracy equal across question sets but varied the amount of expected effort. Experiment 2 varied the expected task accuracy across question sets but held the expected effort constant. Experiment 3 varied both the expected task accuracy and the expected effort in opposite directions. A fourth experiment was conducted as a control, using the same parameters as Experiment 3, but with skip trials that were clearly impossible to read (7% visibility; see bottom panel of Figure 1), which completely obscured the TRT target word.

Fourth, participants were asked to make subjective ratings at the end of each set of multiple-choice questions (i.e., twice per experiment). Subjective questions included ratings of either performance, effort, or work. The surveys regarding "performance" asked, "Please use the slider to rate how well you did on the task you just finished. The scale ranges from 0 to 100." Anchors of "none correct," "some correct," and "all correct" were displayed above the slider bar. This question was based on similar questions used in previous investigations (e.g., Cienkowski & Speaks, 2000; Cox et al., 1991). The surveys regarding "effort" asked, "Please use the slider to rate how much effort it took for you to complete the task you just finished. The scale ranges from 0 to 100." Anchors of "no effort," "some effort," and "extreme effort" were displayed above the slider bar. This question was based on similar questions used in previous investigations (e.g., Brons et al., 2013; Luts et al., 2010). As a synonym for "effort," a survey using the word "work" was included, with instructions designed to call attention to the difference between performance (the potential heuristic substitution under investigation) and the desired construct. The surveys regarding "work" asked, "On average, how hard did you have to work to answer the questions? Keep in mind, how much work a question took does not depend on whether you got the question right or wrong; we are only interested in how much mental work it took to answer the questions. Please use the slider to make your rating. The scale ranges from 0 to 100." Anchors of "no work," "some work," and "extreme work" were displayed above the slider bar. This question was based on one used in previous investigations (e.g., Picou et al., 2017; Picou & Ricketts, 2018). All ratings were made by moving a digital slider with numerical limits of 0–100, in increments of 1. The numerical value of the slider position was visible to participants and was updated in real time as the slider was moved. Written anchors were positioned at the ends and middle of the slider bar. Test items and selections remained visible to participants as they made their ratings. No feedback was provided regarding the number of items correctly answered. To restrict participant responses to a single rating dimension, any single cohort of participants answered a single subjective question. Order of set presentation was counterbalanced within a cohort.

Procedures

During data collection, a survey with a single subjective question (e.g., perceived performance) was posted at a time. Once the predetermined number of participants completed the survey, a new survey was opened for additional participants with a different subjective question (e.g., effort). Within the Mturk website, participants read general information about the study, including expected time course, inclusion criteria, and payment. The study description indicated that this was a test of language proficiency containing vocabulary matching questions. If a potential participant was interested in participating, he or she clicked on a link that opened a new browser window and was redirected to the survey. Study data were collected and managed using Research Electronic Data Capture (REDCap), an electronic data capture tool hosted at Vanderbilt University (Harris et al., 2009). REDCap is a secure, web-based application designed to support data capture for research studies. Upon completion, the participant entered a unique identifying code in the Mturk window and submitted a certification that the survey was completed. If a survey was not completed within 20 min, the session expired and a participant was not compensated. Testing took approximately 5 min for each participant. Mturk workers reported their anonymous Mturk worker identification numbers, which allowed for exclusion of participants who took the survey more than once. No participants repeated the survey. Experiments were conducted sequentially. Within an experiment, the order of survey posting was randomized.

Experiment 1

The purpose of this experiment was to evaluate subjective ratings when expected task accuracy was held stable and expected effort varied between the two sets of multiple-choice questions (see Table 1). Expected effort was increased by reducing the visibility of the target words in Set 2 relative to Set 1 (46% and 70%, respectively). Expected task accuracy was manipulated by including eight stimuli in each set whose visibility did not limit accuracy and two skip trials that were impossible to read (28% visible). The skip trials were included to reduce task performance out of the ceiling. The hypothesis was that participant ratings of performance, effort, and work would reflect the experimental condition, because keeping task accuracy constant across sets effectively removes accuracy as an available heuristic response strategy. Without the familiar percent scale to rely on for rating effort, participants would have to engage an effortful strategy and perhaps make an unbiased rating based on cues related to the actual prompt.

Results and Discussion

Figure 2 displays mean task accuracy and subjective ratings for Experiment 1. Accuracy scores and subjective ratings were analyzed separately, each with a mixed-model analysis of variance (ANOVA) with one within-subject factor, Set (Set 1 and Set 2), and one between-subject factor, Question (performance, effort, and work). Regarding task accuracy, results revealed no significant main effects or interactions, indicating that neither question nor condition affected task accuracy ($p > .10$). Moreover, as expected, accuracy was 83.99% averaged across all conditions.

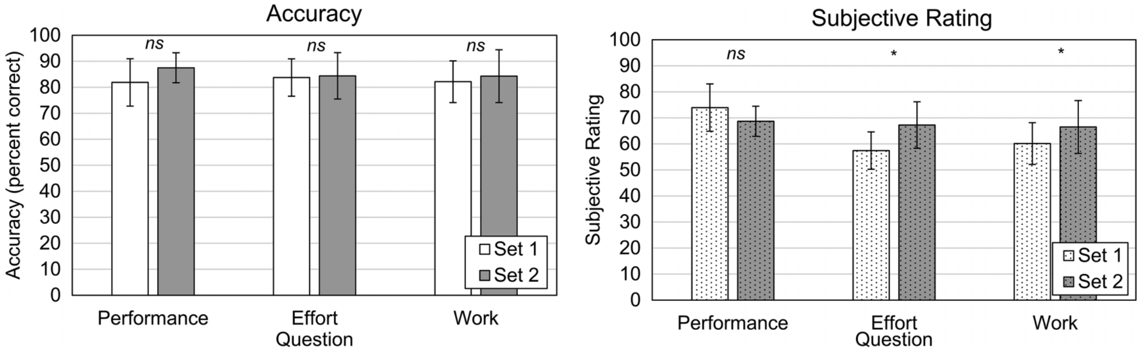
Regarding subjective ratings, results revealed no significant main effect of Set ($p > .05$) or Question ($p > .2$). There was a significant Set × Question interaction ($F[2, 43] = 5.74$, $p < .01$, $\eta_p^2 = .21$), indicating that the effect of Set varied based on the subjective question used to elicit the subjective rating. To explore the interaction, separate ANOVAs with a single within-subject factor (Set) were conducted for ratings in response to each question. Results revealed no effect of Set for ratings of performance ($F[1, 15] = 2.87$, $p > .10$, $\eta_p^2 = .16$, M difference = −5.25, 95% CI [−11.86, 1.36]). However, results revealed that ratings of effort ($F[1, 15] = 6.13$, $p < .05$, $\eta_p^2 = .29$, M difference = 9.81, 95% CI [−1.36, 18.26]) and work ($F[1, 13] = 5.56$, $p < .05$, $\eta_p^2 = .30$, M difference = 6.36, 95% CI [0.53, 12.18]) were significantly higher in Set 2 than in Set 1.

These results demonstrate that, as hypothesized, there were no significant differences in ratings of performance between Set 1 and Set 2 and there was no significant difference in actual task accuracy. Also consistent with expectations, there was a significant difference in the subjective reports of "effort" and "work" when more visual masking was present and visibility was reduced. These data demonstrate that increased masking increased reported mental effort. Because performance was stable across sets, perceived performance did not bias subjective ratings of effort because performance was not a ready alternative cue for making ratings. Importantly, these data also suggest that the experimental manipulations of performance and effort were successful.

Experiment 2

The purpose of this experiment was to evaluate subjective ratings when expected task accuracy varied and expected effort was stable across the two sets of multiple-choice questions (i.e., the opposite conditions of Experiment 1; see Table 1). As in Experiment 1, task accuracy

Figure 2. Mean question accuracy (left panel) and subjective rating (right panel) for Experiment 1. Set 1 (white bars) included eight words that were mostly visible (little effort) and two skip trials (28% visible). Set 2 (gray bars) included eight words that were moderately masked (moderate effort) and two skip trials (28% visible). Error bars are ±1 *SD* from the mean. *ns* denotes nonsignificant differences. *$p < .05$.



was manipulated by including impossible-to-read skip trials (i.e., 28% visible). There were more skip trials in Set 2. Effort was expected to be similar across sets because an equal ratio of low and moderately masked stimuli was included in both sets. If response biases do not play a large role in the subjective ratings under these conditions, it would be expected that ratings of performance would be lower in Set 2, but ratings of effort would be stable across sets. However, if response biases do influence subjective ratings of effort, it would be expected that participants (a) rate performance lower in Set 2 than Set 1, (b) rate effort higher in Set 2 than Set 1, and (c) rate work comparably in Sets 1 and 2, if alerting participants to the potential for bias counteracted the use of heuristics.

Results and Discussion

Figure 3 displays mean task accuracy and subjective ratings, which were analyzed separately as in Experiment 1. Analysis of task accuracy scores revealed a significant main effect of Set, $F(1, 68) = 111.40$, $p < .001$, $\eta_p^2 = .62$, M difference = −21.63, 95% CI [−25.72, −17.54]. There was no significant effect of Question or Set × Question interaction ($p > .05$). These data indicate that participants answered fewer questions correctly in Set 2 than Set 1, but the effect of set was independent of questions used to elicit subjective ratings.

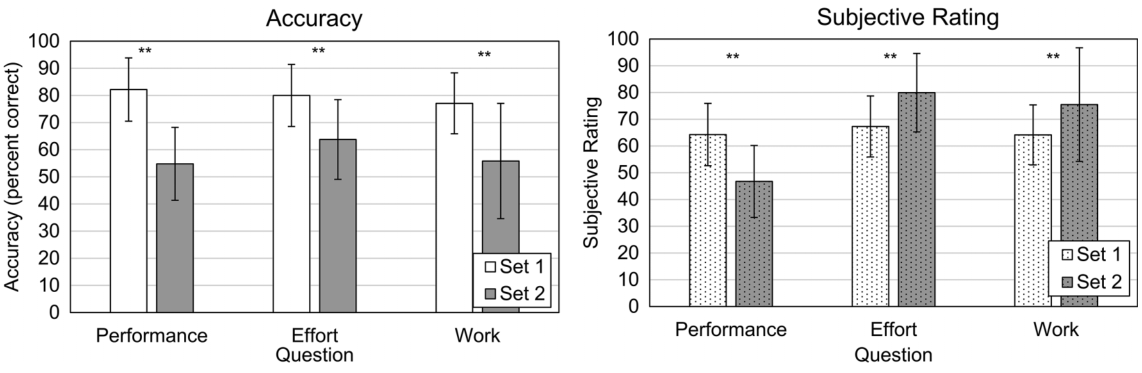
Analysis of subjective ratings revealed no significant main effect of Set ($p > .2$). There was a significant effect of Question, $F(2, 68) = 10.30$, $p < .001$, $\eta_p^2 = .23$, and a significant Set × Question interaction, $F(2, 68) = 20.45$, $p < .001$, $\eta_p^2 = .38$, indicating that the effect of set varied based on the subjective question used to elicit rating. Follow-up testing for each Question revealed significant main effects of Set for ratings of performance, $F(1, 22) = 13.28$, $p < .001$, $\eta_p^2 = .38$, M difference = −17.52, 95% CI [−27.49, −7.55]; ratings of effort, $F(1, 23) = 17.77$, $p < .001$, $\eta_p^2 = .44$, M difference = 12.63, 95% CI [6.43,18.82]; and ratings of work, $F(1, 23) = 12.05$, $p < .01$, $\eta_p^2 = .34$, M difference = 11.33, 95% CI [4.58, 18.09].

These data show that ratings of performance were lower in Set 2, whereas ratings of effort and work were higher in Set 2. As hypothesized, participants were sensitive to the difference in task accuracy across sets, and despite constant expected effort across both sets, subjective ratings of effort increased when performance decreased, strongly suggesting that participants substituted a question concerning performance for the more computationally expensive question regarding mental effort. It seems participants engaged in attribute substitution, basing their ratings of effort on their accuracy when percent correct differed across sets. Counter to expectations, subjective ratings of work also varied with performance, despite the overt wording of the prompt stating that work was separate from performance. As we do not believe that participants intentionally ignore subjective prompts, the change in reported work between sets is interpreted as a testament to the automatic nature of heuristic strategies, as has been suggested by others (e.g., Kahneman & Tversky, 1973).

Experiment 3

The purpose of this experiment was to evaluate subjective ratings when performance and effort were expected to vary in opposite directions (see Table 1). It was expected that, because rating mental effort is difficult, participants would employ an attribute substitution and rate their mental effort according to their performance and not according to the mental effort required to complete the sets. If participants do not disassociate performance from effort when both vary and are instead prone to response bias, it would be expected that ratings of performance would be higher in Set 2 (higher expected performance and effort) and ratings of effort would be higher in Set 1 (lower expected performance and effort). If explicitly directing participants away from the readily available percent correct heuristic during the rating task could overcome the response bias, it would be expected that ratings of work would be higher in Set 2. If heuristic use does not bias ratings of mental effort under

Figure 3. Mean question accuracy (left panel) and subjective rating (right panel) for Experiment 2. Set 1 (white bars) included four words that were mostly visible (little effort), four words that were moderately masked (moderate effort), and two skip trials (28% visible). Set 2 included two words that were mostly visible (little effort), two words that were moderately masked (moderate effort), and six skip trials (28% visible). Error bars are ±1 *SD* from the mean. $^{**}p < .01$.



these conditions, ratings of performance and effort would be higher in Set 2 (higher expected performance and effort).

Results and Discussion

Figure 4 displays mean task accuracy and subjective ratings, which were analyzed separately as in Experiments 1 and 2. Analysis of accuracy scores revealed a nonsignificant main effect of Question ($p > .4$); a significant main effect of Set, $F(1, 67) = 118.73$, $p < .001$, $\eta_p^2 = .62$; and a significant Set × Question interaction, $F(2, 67) = 3.512$, $p < .05$, $\eta_p^2 = .10$. Separate ANOVAs with a single within-subject factor were conducted for accuracy for each question. Results revealed significant main effects of Set for accuracy when the subjective question asked about performance ($F[1, 27] = 96.51$, $p < .001$, $\eta_p^2 = .78$, *M* difference = 27.14, 95% CI [21.47, 32.81]), effort ($F[1, 23] = 18.53$, $p < .001$, $\eta_p^2 = .46$, *M* difference = 17.08, 95% CI [8.87, 25.29]), and work ($F[1, 23] = 58.14$, $p < .001$, $\eta_p^2 = .77$, *M* difference = 17.22, 95% CI [12.46, 21.99]). These data indicate that, for all subjective questions, accuracy was better in Set 2 than in Set 1, although the difference was larger for participants who answered the subjective question about performance.

Analysis of subjective ratings revealed a nonsignificant effect of Set ($p > .8$). There was a significant main effect of Question, $F(2, 67) = 11.67$, $p < .001$, $\eta_p^2 = .26$, and a significant Set × Question interaction, $F(2, 67) = 16.52$, $p < .001$, $\eta_p^2 = .33$, indicating that the effect of Set varied based on the subjective question used to elicit rating. Follow-up testing revealed significant main effects of Set for ratings of performance, $F(1, 27) = 19.37$, $p < .001$, $\eta_p^2 = .42$, *M* difference = 11.39, 95% CI [6.08, 16.71]; ratings of effort, $F(1, 23) = 4.87$, $p < .05$, $\eta_p^2 = .18$, *M* difference = −4.75, 95% CI [−9.20, −0.30]; and ratings of work, $F(1, 17) = 6.27$, $p < .05$, $\eta_p^2 = .27$, *M* difference = −7.61, 95% CI [−14.02, −1.20]. These data show that ratings of performance were higher in Set 2 than in Set 1, but ratings of effort and work were lower in Set 2 than in Set 1.

As hypothesized, the pattern of results once again demonstrated that perceived performance biased subjective ratings of effort and work. That is, participants rated their performance as higher but their effort and work as lower for Set 2 compared with Set 1. This finding is noteworthy because, despite higher performance in Set 2, all TRT words were more heavily masked in Set 2 compared with Set 1 and therefore required more effort. Results from Experiment 1, where performance was stable and effort varied, confirmed that more masking required more effort; however, when performance served as a ready heuristic, it seems the heuristic response strategy dominated. Contrary to initial expectations, the use of the attribute substitution heuristic also seems to have occurred for ratings of work, despite careful question wording intended to guide ratings away from performance.

Experiment 4

Experiment 4 was conducted as a control condition to rule out the possibility that participants exerted effort on skip trials with 28% visibility (considered impossible to read in Experiments 1–3). To achieve this, Experiment 4 had the same parameters as Experiment 3 but used a TRT target word that was only 7% visible (97% masking). This effectively ruled out the possibility of participants working hard to figure out a very challenging condition (i.e., 28% visible stimuli), by utilizing masking that effectively covered the entire TRT word. Given the high similarity between ratings of effort and work in Experiments 1–3, the subjective question related to work was abandoned in Experiment 4.

Results and Discussion

Figure 5 displays mean task accuracy and subjective ratings for Experiment 4. Analysis of accuracy scores revealed a significant main effect of Set, $F(1, 84) = 93.13$, $p < .001$, $\eta_p^2 = .53$. The main effect of Question and the Set × Question interaction were not significant (*p*s = .08 and .44, respectively). Analysis of subjective ratings revealed a nonsignificant main effect of Set ($p = .11$). There were significant effects of Question, $F(1, 84) = 10.80$, $p < .01$, $\eta_p^2 = .11$, and a significant Set × Question interaction,

Figure 4. Mean question accuracy (left panel) and subjective rating (right panel) for Experiment 3. Set 1 (white bars) included four words that were largely visible (little effort) and six skip trials (28% visible). Set 2 included eight words that were moderately masked (moderate effort) and two skip trials (28% visible). Error bars are ±1 *SD* from the mean. *$p < .05$. **$p < .01$.

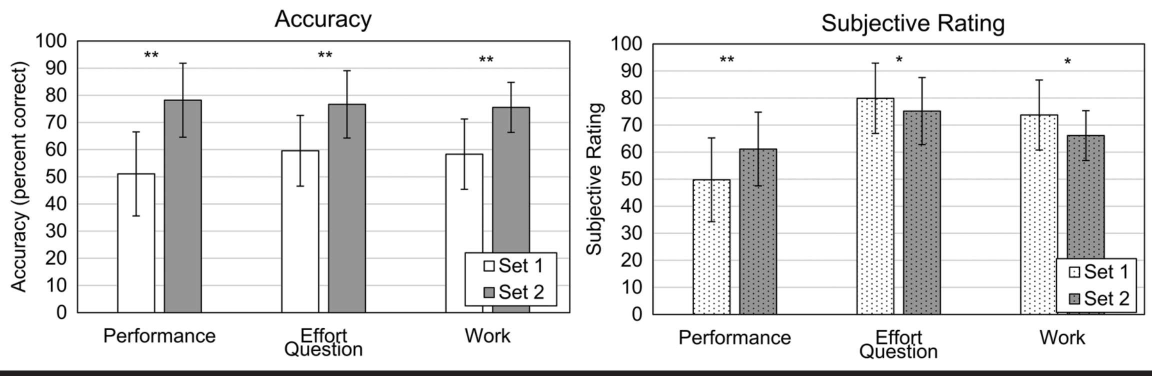
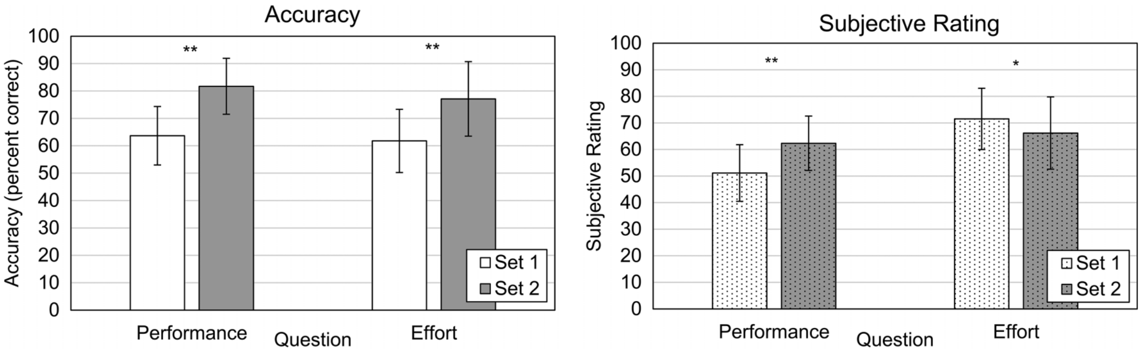


Figure 5. Mean question accuracy (left panel) and subjective rating (right panel) for Experiment 4. Set 1 (white bars), the easy condition, included four words that were largely visible (little effort) and six skip trials (7% visible). Set 2 (gray bars) included eight words that were moderately masked (moderate effort) and two skip trials (7% visible). Error bars are ±1 *SD* from the mean. **p* < .05. ***p* < .01.



$F(1, 84) = 21.08$, $p < .001$, $\eta_p^2 = .20$, indicating that the effect of Set varied based on the subjective question used to elicit rating. Follow-up testing revealed significant main effects of Set for ratings of performance, $F(1, 40) = 16.06$, $p < .001$, $\eta_p^2 = .29$, *M* difference = 11.22, 95% CI [5.56, 16.88], and ratings of effort, $F(1, 44) = 5.33$, $p < .05$, $\eta_p^2 = .11$, *M* difference = −4.33, 95% CI [−9.99, −0.68]. These data show that ratings of performance were higher in Set 2 than in Set 1, but ratings of effort and work were lower in Set 2 than in Set 1. The pattern of results in Experiments 3 and 4 is the same, suggesting that increasing the masking level did not affect the results. The data from Experiment 4 support the hypothesis that participants guessed on the skip trials when visibility was either 28% or 7%.

General Discussion

This study investigated the possibility that study participants substitute the question "How did I perform?" when faced with the more difficult question of "How much effort did that take?". The potential for task accuracy to bias subjective ratings of mental effort was investigated using experimental conditions where accuracy either would or would not be available as a heuristic attribute substitution. The combined results from the four experiments revealed that, when task accuracy was a ready cue, perceived performance biased ratings of mental effort (as in Experiments 2–4). Conversely, when performance was removed as a ready substitute, as in Experiment 1, ratings of subjective effort corresponded to expected mental effort.

These results suggest that the apparent face validity of directly asking participants to rate their subjective experience does indeed rely on participants' ability to provide an unbiased response. It seems valid responses can be difficult to elicit depending on the parameter under investigation (e.g., mental effort). These results are consistent with the findings of Zekveld and Kramer (2014), who asked participants to indicate how often they had given up after performing both TRT (visual) and SRT (auditory) tests with various levels of masking. They found that participants reported giving up more often in the most difficult conditions yet also reported high subjective effort in those same conditions. Overall, smaller pupil dilation in the most difficult conditions, consistent with disengagement and low processing load, supported the subjective report of giving up when trial conditions became too difficult but is counter to expectations for subjective reports of high effort. It seems quite plausible that participants provided an unbiased answer regarding how often they gave up, which is a straightforward question quantified with the familiar percent scale, but substituted the difficult question about effort with the easier question "How well did I perform?". Behavioral performance data support this interpretation, as they revealed floor performance in the most difficult conditions.

These results, taken together with the current study findings, support the notion that subjective ratings of listening effort disagree with behavioral and physiological measures not because participant report is sensitive to some unknown aspect of listening effort (e.g., Hornsby, 2013; Mackersie et al., 2015) but because participants are unable to answer the effort question and so answer a related yet different question altogether. The use of heuristic judgments helps explain why subjective report often differs from behavioral and physiological results (e.g., Desjardins & Doherty, 2013; Gosselin & Gagné, 2011b; Hicks & Tharpe, 2002; Hornsby, 2013; Larsby et al., 2005; Mackersie & Cones, 2011; Mackersie et al., 2015; Miyake, 2001; Picou et al., 2017; Wilson & Sasse, 2001; Yeh & Wickens, 1988; Zekveld et al., 2010, 2011). It seems clear that more research is warranted to identify and account for biases in subjective measures of effort, especially when subjective ratings are the only effort-related study outcome. This is not to imply that subject ratings have no place in measuring effort. Subjective ratings, whether biased or not, can still provide a useful pattern of results that can offer insight into participant decision processes, preferences, and future actions.

Potential Strategies for Handling Response Bias

There are several potential strategies for handling response biases related to heuristic use, including changing the wording of the question, training participants, and using subjective ratings only in certain situations. First, it might be possible to limit the effects of heuristic bias on subjective ratings of effort by changing the question, for example, by querying participants about factors associated with effort, such as the desire to give up or improve a mentally challenging situation, rather than explicitly using the word "effort." In the listening effort domain, such questions have been implemented with mixed success (Picou et al., 2017; Picou & Ricketts, 2018). A similar approach would be to draw a respondent's attention to the separate constructs of task accuracy and mental effort. However, this study suggests that changing the wording, such as the modified instructions for the "work" question, is insufficient to avoid judgment heuristics. This is consistent with findings that show that introducing proper scoring rules and drawing attention to heuristic strategies can be insufficient to obviate various heuristic biases (e.g., Kahneman, 2011; Kahneman & Tversky, 1973; Winkler, 1967).

Second, Kahneman and Frederick (2002) suggest that training and experience could be tools to avoid judgment heuristics. They report that errors related to one type of heuristic were reduced when statistically sophisticated individuals were presented with a question modified to accentuate a statistical pitfall in formulating a response; however, statistically naïve individuals continued to use the heuristic strategy. This raises the question of whether increased statistical knowledge or other training might lead to less biased responses.

Winkler (1967) investigated the effect of structured training on the ability of participants to quantify their judgments during a gambling task, which is prone to heuristic use. Winkler found that graduate students with varying degrees of statistical competence (ranging from an introductory course to a PhD in statistics) were eventually able to plot various probability distributions using a written questionnaire, an answer sheet, and graph paper. After training, some participants were able to develop probability distributions that aided their performance in a gambling task. What is most notable is that participants required feedback to improve. Feedback was possible due to the right-or-wrong nature of the simple probability tasks used in their study. However, without a gold standard measure of listening effort, researchers cannot provide meaningful feedback to study participants, which may be necessary for the average individual to learn to produce unbiased ratings of effort.

Finally, from the current results, it seems that overcoming this particular bias may require abolishing it as a basis for judgment, as in Experiment 1 (where task accuracy was equated and subjective ratings of effort differed with the amount of masking as expected). That is, one strategy for circumventing heuristic biases in subjective ratings of effort might be to evaluate subjective effort when task accuracy is stable across conditions (e.g., ceiling performance). It is also noteworthy that Zekveld and Kramer (2014) only observed incongruous results between pupil responses and subjective ratings of effort when performance was low. Their finding is consistent with the current study, which also showed that the heuristic bias was most evident when accuracy was low. Thus, another approach may be to measure effort when performance is high across conditions. Further work is needed to establish the conditions under which subjective ratings of effort are not biased by heuristic questions. As these scenarios are identified, they may prove beneficial to include in the assessment of construct validity of subjective questionnaires of complex psychophysiological phenomena, including listening effort.

Choice of Text Masking Levels

Effort and performance were controlled in this study by varying the amount of unmasked text available. The study conclusions necessarily rely on several assumptions related to experimental design, namely, (a) reduced visibility from 70% to 46% increased effort without affecting task accuracy, and (b) participants exerted little mental effort on skip trials (28% and 7% visibility). Support for the first assumption comes from the measured task accuracy. Collapsed across all conditions and experiments, task accuracy was near ceiling for the 70% and 46% visibility conditions, 93.5% (*SD* = 8.4%) and 91.0% (*SD* = 8.2%), respectively. These data indicate that participants successfully answered questions with both degrees of visibility and that the increased masking was slightly more difficult.

Support that the masking manipulations successfully affected effort comes from the data in Experiment 1, where effort ratings presumably reflect unbiased subjective ratings of effort because task accuracy was not different between sets (see Figure 2). Conversely, ratings of effort and work were significantly higher in Set 2 (where stimuli were 46% visible) than in Set 1 (where stimuli were 70% visible). It is also important to acknowledge, however, that the actual cue used for subjective ratings of effort in Experiment 1 is unknown and may in theory differ from a predefined concept of "effort." Further work is necessary to validate that the subjective ratings of effort under conditions such as those in Experiment 1 are indeed unbiased indicators of mental effort.

Regarding the second assumption, that participants did not exert effort on the impossible-to-read skip trials (7% visible), it has been reported in the literature that people disengage or exert little effort in exceptionally difficult situations (e.g., Wu et al., 2016; Zekveld & Kramer, 2014). Additional supporting evidence stems from at least three sources in this study: results of Experiment 4, average survey completion time, and the distribution of correct responses.

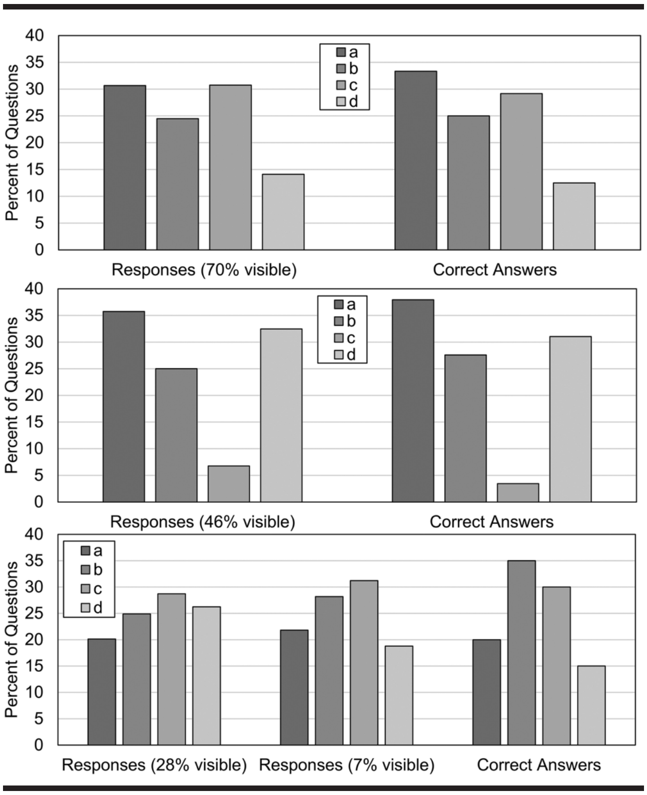
First, for Experiments 1–3, the skip trials included target words that were 28% visible. Inspection of the third panel of Figure 1 reveals that some of the word was visible.

Thus, it is possible that some participants exerted effort to reconstruct the masked word and match it to an appropriate synonym. Thus, one might expect higher effort in any set that included more impossible trials. However, Experiment 4 replicated Experiment 3, but with clearly impossible skip trials with only 7% of the word visible (see bottom panel of Figure 1). Given the impossibility of this task, it would be unlikely that a participant could reconstruct the word and successfully identify the synonym. The pattern of results in Experiments 3 and 4 was identical, suggesting that participants indeed responded arbitrarily to the skip trials in all experiments.

Second, the average survey duration was 5.35 min, including the demographic questions, catch trials, 20 multiple-choice questions, and two subjective ratings. This duration does not allow sufficient time to employ effortful strategies in trying to match impossible-to-read words to a random group of potential synonyms. Moreover, participants were compensated the same rate regardless of their performance on the tasks. There would be no external incentive for participants to exert high effort to decode heavily masked words.

Finally, the distribution of correct responses for each level of masking supports low-effort, random guessing on skip trials. Specifically, for the stimuli where performance was nearly perfect (70% and 46% visibility), the distributions of actual responses and correct responses were quite similar (see top and middle panels of Figure 6). Conversely, the bottom panel of Figure 6 demonstrates a relatively flat distribution of answers for skip trials, where task accuracy was poor (36.4% [*SD* = 11.8%] and 42.9% [*SD* = 9.7%] for the 28% and 7% visible stimuli, respectively). Although the distributions are relatively flat, there is a slight increased likelihood of “b” or “c” responses for both levels of masking, which is consistent with previous work demonstrating that participants tend to favor “b” and “c” response options when guessing randomly on four response-option, multiple-choice tests (Attali & Bar-Hillel, 2003), particularly for low-stakes computer tests (Wise, 2006). Indeed, Wise (2006) reported that 59% of responses due to random guessing were either “b” or “c” options. In this study, 59.2% of responses were either “b” or “c,” confirming random guessing on skip trials.

Figure 6. Distribution of actual responses and correct answers for each level of masking. Distributions are collapsed across experimental conditions.



Study Limitations

Applicability to Listening Effort

A limitation of the current study’s applicability to listening effort per se is the use of visual stimuli to evaluate mental effort. In a study using pupillometry to quantify cognitive load, Klingner, Tversky, and Hanrahan (2011) showed that visual presentation of stimuli led to lower cognitive load than auditory presentation across a range of tasks (i.e., arithmetic, memory, and vigilance). This finding could suggest that the findings with visual stimuli have limited generalizability for auditory stimuli. However, it is worth noting that, if subjective ratings of effort are based at least in part on cognitive load, the results of an auditory experiment could reveal even more marked results than the visual task elicited here.

Furthermore, there is evidence suggesting that listening effort and the more general construct of mental effort share general linguistic processes. Humes et al. (2007) investigated performance on auditory and visual measures of speech processing and showed that, whereas performance was most similar within a single modality, performance was also similar across modalities (see also Watson, Qiu, Chamberlain, & Li, 1996). Principal component analysis revealed that cognition in the visual domain could account for 25%–50% of the variance in the auditory domain. These results demonstrate that auditory and visual processing of linguistic stimuli is not completely independent but instead interacts to influence some supramodal factors, such as cognitive processing. Although the current study did not directly investigate listening effort, it did address the shared cognitive effort inherent in linguistic tasks regardless of modality, using existing visual analog stimuli shown to correlate with results from a common auditory speech test. Nevertheless, future studies are needed to verify whether these results replicate with auditory stimuli. Although the objective of recruiting over 250 participants and the use of crowdsourcing precluded the use of auditory stimuli in this study, subsequent work might focus on a particular combination of task accuracy and effort with a smaller sample size more amenable to typical listening effort procedures in the laboratory. If these study findings can be replicated, for example, using a dual-task paradigm and auditory stimuli, the case for heuristic judgments influencing listening effort will be strengthened.

Design Limitations

A design limitation of the current study stems from our use of published psychometric data in choosing masking levels rather than measuring TRT thresholds for each individual participant. Subsequently, we cannot offer further insight into the range of text reception abilities present in the study population and cannot confirm the shape of the psychometric function as similar to that in Zekveld et al. (2007). However, this study did use masking values well above and below those reported in the literature for comparable performance levels to help account for variability in TRT ability across participants. TRT data reported here for single-word stimuli agreed well with the published data using sentences.

Another potential design limitation is the use of crowdsourcing. Although crowdsourcing has been shown to be a valid and reliable data collection tool used in psychology (Buhrmester et al., 2011; Crump et al., 2013) and audiology (Barber & Lee, 2015; Singh et al., 2015), by posting surveys online, we lose some degree of experimental control. We must trust the integrity of the data because data collection was not directly supervised. Consequently, we instituted a system of data checks and successfully eliminated spurious data from respondents who did not follow instructions (see Table 3). However, the possibility of inauthentic participant responses cannot be fully excluded.

The heterogeneity in the age of participants might also limit the interpretability of the data. Although generally stable across groups, most cohorts ranged in age from young adult (19–21 years) to middle-aged adult (50–68 years), with an average age of approximately 35 years across all experiments. The interaction between heuristic use and age has yet to be explored, but previous results do suggest that age affects mental effort (e.g., Deaton & Parasuraman, 1993; Tomporowski, 2003). In the auditory domain, older adults exert more listening effort than their younger peers, even when both groups have normal hearing (e.g., Gosselin & Gagné, 2011a, 2011b). Thus, the exerted effort of the older participants may have been different than that of the younger participants in this study, which could have affected the subjective ratings of effort or work. Differences in participant age may also have resulted in discrepancies in computer literacy, which could have influenced the subjective effort involved in completing the task. Future studies are warranted to evaluate systematically the interaction between age and heuristic use for subjective ratings of effort.

Future Directions

Behavioral validation of effort was not possible in the current study design. Future studies using online methods should directly query participants concerning their response strategies for different conditions to obtain a better understanding of the cognitive activity employed during the task. For instance, insight into participants' decision to give up could elucidate the extent to which they actually exerted effort, such as in the study by Zekveld and Kramer (2014), who found smaller pupil dilation when participants indicated that they had given up. Similarly, future laboratory investigations using behavioral indications of effort, such as response times or secondary task performance, would also be useful in determining the extent to which a rating of subjective effort is biased.

Future work should also investigate other heuristic strategies as they relate to subjective report of effort. Although this study focused on substituting an easier question for a difficult one, there are many different judgment heuristics (for a review, see Kahneman, 2003, 2011). For instance, it has been shown that judgments can be influenced by affective responses to information (e.g., Slovic, Finucane, Peters, & MacGregor, 2007; Winkielman, Zajonc, & Schwarz, 1997). Applied to the current study, the negative feelings associated with poor performance in conditions with more skip trials (or the converse) might have contributed to ratings of effort. Another potential bias that could affect subjective ratings of effort or performance, but was not controlled in this series of experiments, is the weight a participant applies to each trial type (i.e., low effort, moderate effort, and skip). Although participants may have reflected upon the relative frequency of each trial type to inform their rating of effort, it has been shown that people are poor at integrating such information (e.g., Dawes, 1979). For example, difficult trials might be more heavily weighted if they are encountered within the context of many easier trials. Studies that explore the relative influence of these and other heuristic phenomena are needed to understand fully the information provided by subjective measures of effort and how they contribute to our understanding of listening effort.

Conclusion

In summary, the current study supports the use of judgment heuristics as a possible explanation for the discrepancy between subjective ratings and behavioral and physiological measures of listening effort. It seems the complexities of assigning value to an unfamiliar and difficult-to-quantify concept such as listening effort deserve further consideration in the literature, which already successfully employs behavioral and physiological measures capable of indexing the resource allocation associated with effortful performance. Finding the conditions and rating domains that lead to agreement among the three common measures of listening effort discussed here (subjective, behavioral, and physiological) will help deepen current understanding of the nature of the psychophysiological impact of this complex phenomenon. Although further work is necessary before specific recommendations can be made, the results of this study suggest that (a) ratings of effort are likely based on simpler, substituted questions that may or may not reflect the functional definition of effort and (b) instructions intended to avoid the use of a possible heuristic strategy are insufficient to elicit an unbiased response when rating effort. Future studies using subjective ratings of complex psychophysiological phenomena, such as listening effort, would do well to include study-specific questions that probe the nature of the underlying processes participants use in determining their ratings. Gopher and Braune (1984, p. 520) succinctly capture the nature of the challenge ahead: "Human subjects appear to have no difficulty in assigning numerical values to their experience. However, the experimenter has the burden of selecting the appropriate dimensions for rating."

Acknowledgments

This project was funded in part by Sonova AG (awarded to Erin Picou), and the development of REDCap was supported by the Vanderbilt CTSA Grant UL1 TR000445 from NCATS/NIH. The article contents are the authors' sole responsibility and do not necessarily represent official National Institutes of Health views. The authors would like to thank Todd Ricketts, Rocky Nachalo, and Jake Cantor for their insight and suggestions throughout the project.

References

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*(2), 109–128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Barber, S. J., & Lee, S. R. (2015). Stereotype threat lowers older adults' self-reported hearing abilities. *Gerontology, 62*(1), 81–85. https://doi.org/10.1159/000439349

Broadbent, D. E. (1958). *Perception and communication*. Amsterdam, the Netherlands: Elsevier Scientific.

Brons, I., Houben, R., & Dreschler, W. A. (2013). Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear and Hearing, 34*(1), 29–41. https://doi.org/10.1097/AUD.0b013e31825f299f

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. https://doi.org/10.1037/e527772014-223

Carrow-Woolfolk, E. (1999). *CASL: Comprehensive Assessment of Spoken Language*. Circle Pines, MN: AGS.

Cienkowski, K. M., & Speaks, C. (2000). Subjective vs. objective intelligibility of sentences in listeners with hearing loss. *Journal of Speech, Language, and Hearing Research, 43*(5), 1205–1210. https://doi.org/10.1044/jslhr.4305.1205

Cox, R. M., Alexander, G. C., & Rivera, I. M. (1991). Comparison of objective and subjective measures of speech intelligibility in elderly hearing-impaired listeners. *Journal of Speech and Hearing Research, 34*(4), 904–915. https://doi.org/10.1044/jshr.3404.904

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571–582. https://doi.org/10.1037//0003-066x.34.7.571

Deaton, J. E., & Parasuraman, R. (1993). Sensory and cognitive vigilance: Effects of age on performance and subjective workload. *Human Performance, 6*(1), 71–97. https://doi.org/10.1207/s15327043hup0601_4

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and Hearing, 34*(3), 261–272. https://doi.org/10.1097/aud.0b013e31826d0ba4

Engelhardt, P., Ferreira, F., & Patsenko, E. (2009). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology, 63*(4), 639–645. https://doi.org/10.1080/17470210903469864

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research, 53*(1), 18–33. https://doi.org/10.1044/1092-4388(2009/08-0140)

Gagné, J.-P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing, 21,* 1–25. https://doi.org/10.1177/2331216516687287

Gatehouse, S., & Noble, W. (2004). The speech, spatial and qualities of hearing scale (SSQ). *International Journal of Audiology, 43*(2), 85–99. https://doi.org/10.1080/14992020400050014

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors, 26*(5), 519–532. https://doi.org/10.1177/001872088402600504

Gosselin, P. A., & Gagné, J.-P. (2011a). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *International Journal of Audiology, 50*(11), 786–792. https://doi.org/10.3109/14992027.2011.599870

Gosselin, P. A., & Gagné, J.-P. (2011b). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54*(3), 944–958. https://doi.org/10.1044/1092-4388(2010/10-0069)

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52,* 139–183. https://doi.org/10.1016/s0166-4115(08)62386-9

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research, 45*(3), 573–584. https://doi.org/10.1044/1092-4388(2002/046)

Hornsby, B. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing, 34,* 523–534. https://doi.org/10.1097/aud.0b013e31828003d8

Humes, L. E., Burk, M. H., Coughlin, M. P., Busey, T. A., & Strauser, L. E. (2007). Auditory speech recognition and visual text recognition in younger and older adults: Similarities and differences between modalities and the effects of presentation rate. *Journal of Speech, Language, and Hearing Research, 50*(2), 283–303. https://doi.org/10.1044/1092-4388(2007/021)

Johanssen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., & Wickens, C. (1979). Final report of experimental psychology group. In N. Moray (Ed.), *Mental workload* (pp. 101–114). Boston, MA: Springer.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93,* 1449–1475. https://doi.org/10.1257/000282803322655392

Kahneman, D. (2011). *Thinking, fast and slow*. Basingstoke, United Kingdom: Macmillan.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases* (Vol. 49, pp. 49–81). Cambridge, MA: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48*(3), 323–332. https://doi.org/10.1111/j.1469-8986.2010.01069.x

Kramer, S. E., Zekveld, A. A., & Houtgast, T. (2009). Measuring cognitive factors in speech comprehension: The value of using the text reception threshold test as a visual equivalent of the SRT test. *Scandinavian Journal of Psychology, 50*(5), 507–515. https://doi.org/10.1111/j.1467-9450.2009.00747.x

Larsby, B., & Arlinger, S. (1994). Speech recognition and just-follow-conversation tasks for normal-hearing and hearing-impaired listeners with different maskers. *Audiology, 33*(3), 165–176. https://doi.org/10.3109/00206099409071877

Larsby, B., Hällgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology, 44*(3), 131–143. https://doi.org/10.1080/14992020500057244

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., . . . Froehlich, M. (2010). Multicenter evaluation of signal enhancement algorithms for hearing aids. *The Journal of the Acoustical Society of America, 127,* 1491–1505. https://doi.org/10.1121/1.3299168

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology, 22*(2), 113–122. https://doi.org/10.3766/jaaa.22.2.6

Mackersie, C. L., MacPhee, I. X., & Heldt, E. W. (2015). Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear and Hearing, 36,* 145–154. https://doi.org/10.1097/AUD.0000000000000091

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *The Quarterly Journal of Experimental Psychology: Section A, Human Experimental Psychology, 58*(1), 22–33. https://doi.org/10.1080/02724980443000151

McNair, D., Lorr, M., & Droppleman, L. (1971). *Profile of mood states*. San Diego, CA: Educational and Industrial Testing Service.

Miyake, S. (2001). Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology, 40*(3), 233–238. https://doi.org/10.1016/s0167-8760(00)00191-4

Moore, T. M., Key, A. P., Thelen, A., & Hornsby, B. W. (2017). Neural mechanisms of mental fatigue elicited by sustained auditory processing. *Neuropsychologia, 106,* 371–382. https://doi.org/10.1016/j.neuropsychologia.2017.10.025

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104). Englewood Cliffs, NJ: Prentice Hall.

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage, 55*(2), 713–723. https://doi.org/10.1016/j.neuroimage.2010.12.020

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., . . . Mackersie, C. L. (2016). Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing, 37,* 5S–27S. https://doi.org/10.1097/aud.0000000000000312

Pichora-Fuller, M. K., & Singh, G. (2006). Effects of age on auditory and cognitive processing: Implications for hearing aid fitting and audiologic rehabilitation. *Trends in Amplification, 10*(1), 29–59. https://doi.org/10.1177/108471380601000103

Picou, E. M., Moore, T. M., & Ricketts, T. A. (2017). The effects of directional processing on objective and subjective listening effort. *Journal of Speech, Language, and Hearing Research, 60,* 199–211. https://doi.org/10.1044/2016_jslhr-h-15-0416

Picou, E. M., & Ricketts, T. A. (2014). Increasing motivation changes subjective reports of listening effort and choice of coping strategy. *International Journal of Audiology, 53,* 418–426. https://doi.org/10.3109/14992027.2014.880814

Picou, E. M., & Ricketts, T. A. (2018). The relationship between speech recognition, behavioral listening effort, and subjective ratings. *International Journal of Audiology, 57*(6), 457–467. https://doi.org/10.1080/14992027.2018.1431696

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rabbitt, P. (1966). Recognition: Memory for words correctly heard in noise. *Psychonomic Science, 6*(8), 383–384. https://doi.org/10.3758/bf03330948

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *The Journal of the Acoustical Society of America, 136*(5), 2642–2653. https://doi.org/10.1121/1.4897398

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*(2), 207–222. https://doi.org/10.1037/0033-2909.134.2.207

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*(1), 1–20. https://doi.org/10.1146/annurev.psych.41.1.1

Singh, G., Lau, S.-T., & Pichora-Fuller, M. K. (2015). Social support predicts hearing aid satisfaction. *Ear and Hearing, 36,* 664–676. https://doi.org/10.1097/AUD.0000000000000182

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research, 177*(3), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

Tomporowski, P. D. (2003). Performance and perceptions of workload among young and older adults: Effects of practice during cognitively demanding tasks. *Educational Gerontology, 29*(5), 447–466. https://doi.org/10.1080/713844359

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.21236/ad0767426

Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *The Journal of the Acoustical Society of America, 100*(2), 1153–1162. https://doi.org/10.1121/1.416300

Wilson, G. M., & Sasse, M. A. (2001). *Straight from the heart: Using physiological measurements in the evaluation of media quality*. Paper presented at the Proceedings of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB) Convention, York, England.

Winkielman, P., Zajonc, R. B., & Schwarz, N. (1997). Subliminal affective priming resists attributional interventions.

Cognition & Emotion, 11(4), 433–465. https://doi.org/10.1080/026999397379872

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association, 62*(319), 776–800. https://doi.org/10.2307/2283671

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wu, Y.-H., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric functions of dual-task paradigms for measuring listening effort. *Ear and Hearing, 37*(6), 660–670. https://doi.org/10.1097/aud.0000000000000335

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30*(1), 111–120. https://doi.org/10.1177/001872088803000110

Zekveld, A. A., George, E. L., Kramer, S. E., Goverts, S. T., & Houtgast, T. (2007). The development of the text reception threshold test: A visual analogue of the speech reception threshold test. *Journal of Speech, Language, and Hearing Research, 50*(3), 576–584. https://doi.org/10.1044/1092-4388(2007/040)

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology, 51*(3), 277–284. https://doi.org/10.1111/psyp.12151

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing, 31*(4), 480–490. https://doi.org/10.1097/aud.0b013e3181d4f251

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing, 32*(4), 498–510. https://doi.org/10.1097/aud.0b013e31820512bb

Zekveld, A. A., Pronk, M., Danielsson, H., & Rönnberg, J. (2018). Reading behind the lines: The factors affecting the text reception threshold in hearing aid users. *Journal of Speech, Language, and Hearing Research, 61*(3), 762–775. https://doi.org/10.1044/2017_JSLHR-H-17-0196

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools*. Delft, the Netherlands: Delft University Press.

Appendix

Test Items and Multiple-Choice Options Used for All Surveys

Set 1		Set 2	
Test item	**Options**	**Test item**	**Options**
Street	Ship	Silent	Heavy
	Road		Pretty
	Truck		Sad
	Building		Quiet
Nearer	Slower	Simple	Kind
	First		Easy
	Away		Lazy
	Closer		Different
Coerce	Foretell	Genial	Divine
	Heed		Anxious
	Compel		Fanciful
	Aspire		Cordial
Linger	Remain	Fickle	Erratic
	Initiate		Conscientious
	Need		Arrogant
	Roll		Delicious
Stodgy	Odd	Plunge	Dive
	Dull		Trail
	Soft		Pattern
	Old		Swim
Derive	Obtain	Survey	Picket
	Enlarge		Fertilize
	Disagree		Confess
	Nourish		Examine
Infant	Cradle	Morsel	Package
	Adult		Bite
	Baby		Thread
	Children		Mushroom
Breach	Cleaner	Robust	Strong
	Gap		Breakable
	Bribe		Attractive
	Span		Painful
Raffle	Carnival	Solace	Comfort
	Quilt		Punishment
	Lottery		Deity
	Collection		Tassel
Jostle	Elevate	Thwart	Swindle
	Bump		Hoard
	Snicker		Frustrate
	Stifle		Encourage