On Investigating Inner Experience:

Contrasting Moore & Schwitzgebel and Brouwers et al.

Russell T. Hurlburt

University of Nevada, Las Vegas

In press: *Consciousness & Cognition*

Abstract

Moore and Schwitzgebel (2018) reported that readers report innerly speak the text about 59% of the time. Brouwers et al. (2018) reported that readers innerly speak the text only about 3% of the time. I use this huge discrepancy as a case study to discuss important issues in the investigation of inner experience, including: the training participants; ambiguity in the time being considered; ambiguity in the phenomenon; the desirability of investigating the phenomenon itself, not merely its frequency; bracketing presuppositions; the advantages and disadvantages of large and small sample sizes; influence by the investigator; and the slide from phenomena to reports of phenomena.

Keywords: inner speech; reading; experience sampling; descriptive experience sampling (DES); pristine inner experience

Moore and Schwitzgebel (this issue; hereafter M&S) reported (among other things) that readers report innerly speaking the text about 59% of the time. My colleagues and I (Brouwers et al., 2018; hereafter Brouwers) reported that readers innerly speak the text only about 3% of the time. That is a shockingly huge discrepancy.

I provide here a broad comment on the psychological science of inner experience, using M&S as an exemplary instance of modern psychological science and Brouwers as an exemplary instance of descriptive science. Schwitzgebel and I have debated important issues in consciousness science frequently (Hurlburt & Schwitzgebel, 2007, 2011a,b,c); our confrontations have been constructive and instructive. I hope that we can highlight the M&S / Brouwers discrepant results in the same constructive spirit, encouraging psychological science to consider important issues.

Both M&S and Brouwers discuss several features of experience, but I focus here only on inner speech because the contrast is so stark. Further, I focus primarily on M&S Experiment 2, because that is their largest study.

Background

M&S is a questionnaire and experience-sampling study. Participants used a seven-point Likert scale (1 = *Never*, 4 = *Half of the Time*, 7 = *Always*) to respond to an online questionnaire item "How often do you experience an inner voice when you read? Examples: you hear a voice reading in your head, you hear the characters speaking in your head." Their mean response was 4.6, that is, somewhat more than half the time. Then, using a one-beep experience-sampling technique, M&S beeped participants at a random time between 30 and 90 seconds into reading either a poem, a descriptive story, or a story with dialog. Some of the M&S participants

were recruited from a university subject pool, some from MTurk. Regardless of passage type or recruitment strategy, and compatible with their questionnaire responses, 59ish percent of participants answered *Yes* (they could have selected *No* or *Maybe/Don't Know*) to the experience-sampling question "In the final split-second before the beep, did you experience INNER SPEECH?"

Brouwers is a descriptive experience sampling (DES) study. During four days of natural-environment DES sampling, including a one-hour training ("expositional") interview each sampling day, Brouwers trained participants to try to give complete descriptions of the inner experience that was ongoing when beeped. Then Brouwers beeped those participants eight times while reading complete F. Scott Fitzgerald and Ernest Hemingway short stories; the first beep was perhaps 15 minutes into the reading. In the expositional interview about those beeps samples, participants almost never described innerly speaking the text while they read—only 3% of beeped samples involved such inner speaking.

Brouwers is an exemplary study by DES standards. Let us stipulate that Brouwers was the result of an honest attempt to describe phenomena in as high fidelity as the investigators could muster, but any defense that I make of Brouwers is not intended as a claim that DES is adequate nor especially as a claim that DES is the ultimate method for exploring inner experience.

The issue is that M&S and Brouwers both made claims about the frequency of inner speech that was ongoing in the final split-second before the beep. The M&S claim was about 20 times higher than the Brouwers claim. Understanding the source of that huge discrepancy is of fundamental importance to the psychology of inner experience.

A few years ago in this journal, Chris Heavey and I provided a theoretical essay (Hurlburt & Heavey, 2015) that contrasted questionnaires and experience sampling with descriptive experience sampling. That essay critiqued questionnaires and non-DES experience sampling, but did not make claims about the potential magnitude of between-method discrepancies. The present commentary shows that that magnitude can be huge.

Pristine inner experience

M&S is, apparently, motivated by an interest in pristine inner experience (using Hurlburt's 2011 understanding of "pristine" as intending everyday, naturally occurring, non-manipulated experience). As do many psychology studies, its opening lines invite readers to consider their own pristine experience: "What sorts of conscious experiences do you have while reading? You are, in fact, reading at this very moment. So think, what are you experiencing right now?" There are two fundamental issues: Does M&S adequately simulate natural everyday (pristine) reading? and Do the M&S participants adequately report their experience while reading?

> Regarding the adequacy of the simulation, this is the M&S introduction to the task:
>
> You will now have a short passage to read.
>
> While reading, you will hear a beep and will then be asked questions about your inner experience.
>
> When you hear the beep, stop reading and reflect on your inner experience in the final split-second before the beep.

Then a passage appears. I think the participant's mission in such a situation is *not* merely to read as they normally would; it is to read while simultaneously trying to figure out what kinds of questions they might be asked about their experience, while wondering how long the passage is,

while wondering why this passage and not some other, while trying to figure out what the experiment is really about, and so on. The beep occurs within 30 to 90 seconds; I think it likely that few of those supposedly extraneous issues have been resolved or put aside.

Regarding the adequacy of the report: When the beep occurs, the passage disappears, replaced with a new screen (which participants have not seen before, and which they now must read and interpret):

> NOTE: It is perfectly fine to answer all YES or all NO, and there is nothing wrong with answers here conflicting with previous responses or initially held beliefs.
>
> In the final split-second before the beep, did you experience INNER SPEECH?
>
> Please choose the appropriate response for each item:
>
> Yes No Maybe / Don't Know
>
> ○ ○ ○

At issue is the degree to which a participant's ticking *Yes* (59% do so) reflects their inner experience that was actually ongoing while reading the passage in the split-second before the beep, that is, prior to the passage disappearance, prior to their engagement in the complex task of unpacking the currently-on-screen (and arguably more salient) instruction. I think it likely that the current unpacking of the instruction involves innerly speaking the instruction, and the ticking of the response reflects the instruction-reading experience as well as the passage-reading experience.

It may seem that "When you hear the beep, stop reading and reflect on your inner experience in the final split-second before the beep" is such a simple, straightforward instruction that participants will be able to describe experience adequately despite such interference. However, our DES practice demonstrates that participants almost never grasp such instructions on the first sampling day. Instead of reporting ongoing experience, they give generalities about

experience or theories about experience; they respond to what they take to be the investigator's expectations; and so on, leading DES investigations to discard first-day reports entirely. It requires, in our DES experience, substantial on-the-job ("iterative") training to help participants "cleave to" experience—to report about experience and nothing else. The M&S reports are of the first-day, not adequately trained, variety.

Brouwers is also motivated by an interest in investigating pristine experience. Regarding the adequacy of the simulation, Brouwers attempted to create a situation where pristine reading experience can be approximated with high fidelity. Brouwers had participants read entire short stories, not context-less excerpts. Before reading they had been involved in four days of natural-environment DES sampling—whatever questions they might have had about the procedure or the task had been long since resolved. To the extent that the experimental reading task itself was strange (even though it was designed to be quite natural), participants had ample opportunity to acclimatize themselves to it and to be carried along by the story—the first beep occurred on the 24th page.

Regarding the adequacy of the report: Long before their first while-reading beep, as the result of the four days of natural-environment sampling, participants were well practiced in the grasp-ongoing-inner-experience task and could undertake it without new instruction. They understood what is meant by ongoing-at-the-beep inner experience and how to grasp it immediately after the beep. When beeped while reading, they did not need to read and interpret a new instruction; they could focus on what was ongoing at the moment of the beep.

Thus it seems to me that M&S is a problematic investigation of pristine experience on two counts: it has not adequately constructed a natural situation, and it has not trained participants to cleave to experience. Note that I critique the particulars of M&S because it is

exemplary: many modern studies are motivated by an interest in pristine inner experience but make little effort to set up a situation where pristine experience can show itself in a natural way and make little effort to train participants to notice it with fidelity.

The final split-second before the beep

Inner experience can change dramatically within fractions of seconds. M&S participants, as is typical in non-DES experience-sampling studies, received no training about timing other than the instruction "reflect on your inner experience in the final split-second before the beep." However, our DES practice demonstrates that that instruction, despite its seeming straightforwardness, is not adequate in limiting participants to the split-second-before-the-beep time frame. On their first sampling day, DES participants report experiences that occurred decidedly *after* that split-second (e.g., "I was startled by the beep"), decidedly *before* that split-second (e.g., "I was feeling grumpy all day"), or having nothing temporal to do with that split-second (e.g., "I always talk to myself"). It requires, in our DES experience, substantial on-the-job ("iterative") training to help participants "cleave to" the split-second before the beep—to report about things that were indeed ongoing at that time.

Brouwers participants, prior to their first day of training sampling, received instruction very similar to M&S "reflect on your inner experience in the final split-second before the beep"). However, that instruction was given with the assumption that it would be the first (itself ineffective) step in iterative training, and that the understanding of what is actually the moment under consideration would be refined across several natural-environment sampling days.

To the extent that a participant reports experiences that were ongoing at some time other than the beep, those experiences are no longer a random sample, instead reflecting

some (probably unspecifiable) motivated preference. I think it likely that M&S participants, much more than Brouwers participants, reported about a mix of beeped and non-beeped moments.

The general principle: Unless some careful (probably necessarily iterative) procedure has been implemented, it is not safe to presume that experience-sampling participants focus on the split-second before the beep.

What is "inner speech"?

M&S participants, as is typical in questionnaire and experience-sampling studies, received no training about the definition of the phenomenon "inner speech" other than the two examples provided in the initial questionnaire ("Examples: you hear a voice reading in your head, you hear the characters speaking in your head.") Even if participants were influenced by those examples (our own DES studies suggest that that is unlikely), those examples should not be expected to narrow a participant's original understanding of what should be called inner speech.

Hurlburt (2011, particularly in the vicinity of p. 60) noted that DES participants, early in sampling, begin a description of an at-the-moment-of-the-beep experience with a locution such as "I was saying to myself that…." The DES interviewer might then inquire, "What exactly were you saying at that time?" The typical participant might not recall what they were saying, at which interviewer might respond, "No problem here. But when you sample next time, if a beep happens to catch you saying something to yourself, we'd like to know what you were saying." That kind of exchange is at the heart of the DES iterative training: The interviewer does *not* discourage participants from reporting inner speech, but

instead encourages them to report it with more detail. It often happens that on subsequent sampling days, such participants no longer report inner speech.

DES finds that on their first sampling day, participants frequently use "I was saying to myself that…." to introduce all manner of non-speaking experiences, including feelings, sensory awarenesses, experientially cognitive but unworded experiences, and so on. A participant's report "I was saying to myself that that guy is a jerk" often reflects a non-speaking experience that would be better described as *I was angry*. A participant's report "I was saying to myself that that phone case is really cool" often reflects a non-speaking experience that would be better described as *I was drawn to the sparkly blueness of the case*.

Brouwers participants received approximately four hours of training in describing their experience aimed at whatever phenomena was ongoing at the moment of the beep (including but not limited to inner speech). That training aimed at completeness, but of course falls short of that ideal. However, inner speech, when it occurs, is among the easiest phenomena for participants to describe and for investigators to recognize, so it seems unlikely that Brouwers substantially undercounted inner speech.

M&S is exemplary. I think psychological science needs to work through the likelihood that questionnaires over-report (or otherwise distort) inner speech and many other inner-experience phenomena.

Frequency vs. phenomena

The M&S design focuses *only* on frequency. It is possible to mine the M&S huge data set for inner speech frequency differences between passage (poetry, descriptive fiction, dialog fiction), or between recruitment strategy (subject pool, MTurk); those differences are small. The M&S

data set can never shed light on the phenomenon of inner speech, including whether inner speaking phenomena differ among the reading passages or between participants.

Brouwers was designed not only to measure frequency but also to explore phenomena; it provides rich descriptions of a variety of while-reading worded phenomena that are not generally contemplated, for example a focus on specific words from the text (but which were not merely inner reading), or words innerly spoken in response to the text (content was related to the text but not of the text itself), or innerly spoken unrelated words (apparently not connected to the text).

The M&S failure to explore the phenomena of pristine inner experience is exemplary: very few modern studies make such effort.

Bracketing presuppositions

Hurlburt, Heavey, and Kelsey (2013) held that presuppositions about inner speech are powerful; for example, many experts (as well as non-experts) believe that inner speech is ubiquitous, whereas DES shows it to occur at modest frequencies.

Commentators frequently dismiss DES results because they do not align with their own self-perceptions or with what they take to be the intuitions of experts. Here, for example, it might seem impossible that inner experience while reading only rarely includes inner speaking of the text. Hurlburt (2011) has shown that many if not most people, including experts (e.g., Hurlburt, 2011, Ch. 6 and 12), are often mistaken about the characteristics of their own inner experience. The bracketing of presuppositions is a necessity at all levels of investigation, from the participants to the investigators to consumers of research.

M&S, as is typical of many psychological studies, makes no effort to bracket presuppositions. Brouwers makes substantial efforts. I think psychological science needs to

work out how to determine whether such efforts are effective and/or how to create more effective methods.

Number of participants

M&S used 1233 participants. Using a large sample is desirable when the errors of measurement are random. However, if a large proportion of the population share the same bias or if an investigation induces the same distortion (as would be the case if presuppositions are important), no sample size (no matter how large) will correct it.

Brouwers used 16 DES participants. In general, using small samples has two main risks: lack of power and the potential for unrepresentativeness. Lack of power (high probability of Type II error) is not a problem in Brouwers because there was no Type II error: the Brouwers / M&S difference (3% vs. 59%) is significant ($\chi^2 = 20.356$, $p < .0001$). Brouwers tried to limit the risk of unrepresentativeness by recruiting a largish sample ($N = 260$) from a diverse population, stratifying them on a questionnaire measure of self-talk frequency, and then randomly selecting the DES participants from the upper and lower quartiles of that questionnaire. The stratification procedure gave a slight edge to those who by questionnaire believed themselves to be high self-talkers. It seems unlikely that this procedure produced participants who would by chance happen to be unusually low in inner speech while reading.

In general, large samples are better than small samples. However, using a large sample in no way guarantees fidelity to phenomena. Furthermore, using large samples guarantees that no individual participant will be investigated with care, which implies that the effect of presuppositions cannot be limited.

Investigator/investigation demands

There is a large risk in Brouwers that the investigators wittingly or unwittingly influenced each other and/or their participants to underplay the frequency of inner speech. The iterative training method does indeed provide ample opportunity for such influence to take place. The Brouwers team recognized this risk, and took many precautions to mitigate it, but there is no data-driven way of demonstrating that they were successful. It is indeed possible that despite their best efforts, the Brouwers team under-reported inner speech.

We have tried in many ways to reduce the impact of such demands on DES results, including, for example, inviting skeptics such as Schwitzgebel to examine in detail the DES procedure (Hurlburt & Schwitzgebel, 2007). Replication by others not related to me would be an important step of the kind that psychological science should make.

Report vs. description of phenomena

M&S follows a template frequently used in consciousness science writing: begin by invoking phenomena but then slide to a more objective stance. The M&S beginning lines invoke phenomena as the target of interest ("What sorts of conscious experiences do you have while reading? You are, in fact, reading at this very moment. So think, what are you experiencing right now?"), but the Results and Discussion characterize only their participants' *reports* (59% of *reports* involve inner speech); that is, the Results and Discussion do *not* claim that their participants actually engage in inner speech 59% of the time. That phenomena-sliding-to-report template is, I think, a misdirection that is unproductive for science. If M&S is conceptualized as a study of reports, then (I think) it would have been better had it begun frankly with a focus on a study of reports (e.g., "What sorts of *things would you say* about your conscious experience while reading? You are, in fact, reading at this very moment. So think, *what would you say about* your experience right now?") Alternatively, if M&S is

conceptualized as a study of phenomena, then it would have been better had it devised a method that has the intention of exploring phenomena. Trying to have it both ways, as their phenomena-sliding-to-report template enacts, leads to one of two undesirable outcomes: the consumer will not notice the slide and therefore (mis)interpret M&S as making a claim about phenomena, or the consumer will accept the implication that actual exploration of phenomena is impossible or unnecessary. Either way, the science of phenomena is not advanced (or, worse, is regressed, dulled into insensitivity).

Brouwers, by contrast, consistently targets phenomena; there is no misdirection. It says: we are interested in phenomena, and we will make a principled effort to apprehend those phenomena in high fidelity. One can easily criticize the DES principles or criticize the Brouwers implementation of them. I welcome such criticism, because in the long run such criticism might lead to a superior set of principles and/or implementation thereof, to a method of higher fidelity than DES.

Overall

It is possible and desirable to explore aspects of reading and other important human activities in non-self-report ways, using reaction time, eye-tracking, fMRI, and so on. However, despite widespread criticism of introspection, psychological science has consistently maintained an interest in the phenomena of inner experience—an interest that I think is well placed. However, psychological science has not adequately worked through the requirements that such an interest imposes (Hurlburt, 2011). Skinner (1953), among others, showed why descriptions about inner experience are likely to be mistaken; Hurlburt and Heavey (2001; Hurlburt, 2011) discussed how that can be mitigated.

I am certainly not in a position to claim that the Brouwers 3% has been shown unequivocally to be a better estimate of inner speech while reading frequency than is the M&S 59%; in fact, other DES studies suggest that reading non-short-story passages may involve a frequency of inner speech somewhat higher than 3% (but not approaching 59%). I am, I believe, in a position to suggest that science would be well served by figuring out how to proceed in the face of such stark discrepancies.

If science is to be interested in the phenomena of pristine inner experience, as I think it importantly is, then science should work out rational ways of investigating phenomena. Counting reports of phenomena is one aspect of such investigation, but careful description of phenomena is of fundamental and arguably prior importance. However, as the science currently stands, there are very few independent investigators who are seeking to describe the phenomena of pristine inner experience. It seems that science should work out ways of encouraging more investigations that implement principled ways (Hurlburt 2011) of describing phenomena with fidelity.

References

Brouwers, V. P., Heavey, C. L., Lapping-Carr, L., Moynihan, S., Kelsey, J., & Hurlburt, R. T. (2018). Pristine inner experience while silent reading: It's *not* silent speaking of the text. *Journal of Consciousness Studies, 25*(3-4), 29-54.

Hurlburt, R. T. (2011). *Investigating pristine inner experience: Moments of truth*. Cambridge, UK: Cambridge University Press.

Hurlburt, R. T., & Heavey, C. L. (2015). Investigating pristine inner experience: Implications for experience sampling and questionnaires. *Consciousness and Cognition, 31,* 148-159.

Hurlburt, R. T., & Heavey, C. L. (2001). Telling what we know: describing inner experience. *Trends in Cognitive Sciences, 5,* 400-403.

Hurlburt, R. T., Heavey, C. L., & Kelsey, J. M. (2013). Toward a phenomenology of inner speaking. *Consciousness and Cognition, 22,* 1477–1494.

Hurlburt, R. T., & Schwitzgebel, E. (2007). *Describing inner experience? Proponent meets skeptic.* Cambridge, MA: MIT Press.

Hurlburt, R. T., & Schwitzgebel, E. (2011a). Little or no experience outside of attention? *Journal of Consciousness Studies, 18*(1), 234–252.

Hurlburt, R. T., & Schwitzgebel, E. (2011b). Methodological pluralism, armchair introspection, and DES as the epistemic tribunal. *Journal of Consciousness Studies, 18*(1), 253–273.

Hurlburt, R. T., & Schwitzgebel, E. (2011c). Presuppositions and background assumptions. *Journal of Consciousness Studies, 18*(1), 206–233.

Moore, A.T., & Schwitzgebel, E. (this issue). The experience of reading. *Consciousness and Cognition.*

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.