Author-final version accepted for publication at the Consciousness and Cognition
Publication detail: Consciousness and Cognition 113 (2023) 103534
https://authors.elsevier.com/a/1hFUm3lcz45EgB

Private Speech Amount Positively Predicts Memory Performance in Young Adults

Xinqi Guo[1], Karen Dobkins[1*]

[1] University of California, San Diego

Corresponding Author:

Karen Dobkins, University of California San Diego, McGill 5117, La Jolla, California, 92037,

United States. Email: kdobkins@ucsd.edu

Declaration of interest: None

Abstract

This study used a card-matching game that relies on visual-spatial working memory to investigate whether the amount one talks out loud to themselves (referred to as private speech) predicts cognitive performance in young adults (n = 118, mean age = 20.13 years). Each participant's performance was measured in two “Private Speech” trials, in which they were instructed to complete the game efficiently, while using private speech as much as they can. Using multilevel modeling, we found that participants performed significantly better on trials for which they produced more private speech. This relationship was not moderated by baseline competency on the task (measured in a condition where participants were not instructed to use, and rarely ever used, private speech). The study shows that the degree to which adults use private speech — when instructed to do so, is associated with cognitive performance, which may have important implications for educational/instructional settings.

Keywords: private speech, task performance, expertise reversal effect, visual-spatial working memory

Private Speech Amount Positively Predicts Memory Performance in Young Adults

Humans possess the unique ability to talk to themselves, and although much of this self-talk is kept silent (referred to as "inner speech"), some of it is in the form of talking out loud (referred to as "private speech" or "thinking out loud"). In his seminal work, Vygotsky (1987) theorized about the emergence, and then submergence, of private speech over the course of development. He proposed that private speech emerges from children's day-to-day social interactions with caregivers and serves a self-regulatory function when the caregivers are not around. Gradually, over the course of development, children switch over to using inner speech, which is considered a more mature form of self-talk. This theory has been substantiated by empirical studies showing that the frequency of private speech peaks during the preschool period, after which it gradually decreases in frequency and/or becomes less audible (Berk, 1986; Winsler et al., 2003).

As might be expected given the prevalence of private speech in children, there exists a substantial literature looking at variables that may be associated with children's use of private speech (reviewed in Alderson-Day et al., 2015; Frauenglass & Diaz, 1985; Winsler, 2009). Much of this work has been correlational in nature, asking whether the amount or type of "spontaneous" (i.e., uninstructed) private speech a child uses correlates with another one of their characteristics/abilities. This correlational approach has been addressed in one of two ways. First are studies that measure private speech usage within a specific setting, and then ask whether that usage correlates with some personality trait or a behavioral ability measured at *another* time/setting. For example, one study in 4- to 7-year-olds reported that children who used more self-regulatory private speech during a manual spatial planning task (Tower of London) also showed more sophisticated abilities in narrating about recent events or their earlier childhood (Al-Namlah et al., 2012). Second are studies that measure private speech usage *while* children are performing a cognitive task, asking whether the amount (or type) of private speech correlates with performance on that task. For example, one study reported that when 3- to 5-year-olds are performing a problem-solving task (using Lego blocks to construct a figure from a presented model), those who used more self-motivational and planning-related private speech during the task showed better performance (Mulvihill et al., 2021). Similarly, Sawyer (2017) tested preschool children's performance on a (toy) fishing activity, and found that performance (number of fish caught) was positively predicted by amount of metacognitive private speech and negatively by motivational private speech.

Although the correlations observed between private speech usage and performance in children are suggestive of a beneficial role of private speech for cognitive tasks, they do not provide conclusive evidence of a *causal* relationship or the direction of that relationship. For this, experimental studies must be conducted, wherein performance is compared between conditions where participants are *instructed* to use private speech vs. conditions where they are either given no instruction (and presumably do not talk out loud) or are explicitly instructed to not talk out loud[1]. The few studies that have adopted an experimental approach with children have shown a beneficial effect of private speech on cognitive tasks, with some studies using a within-subjects

[1] Of course, participants may still be using inner speech under conditions where they are given no instructions or explicitly told not to talk out loud. As such, finding no benefit of talking out loud could occur if participants simply switch between using private speech (when instructed to do so) and inner speech (when not instructed to, or instructed to not, talk out loud), and the two types of self-talk are equally effective.

design (Winsler et al., 2007) and others, a between-subjects design (Fernyhough & Fradley, 2005; Lee, 1999; and see Experiment 2 of Müller et al., 2004).

But what about private speech in *adults*? As noted above, Vygotsky (1987) proposed that it largely disappears by late childhood. More recently, however, Fernyhough (2004) revised Vygotsky's theory by adding a "re-entry" process of private speech in adulthood. This revision was motivated, in part, by evidence showing that, under certain conditions, adults do spontaneously use private speech, for example, during challenging and/or complex cognitive tasks (Alarcón-Rubio et al., 2013; Duncan & Cheyne, 2001; Mulvihill et al., 2021), when learning new manual tasks like crafting lanyards (Soskin & John, 1963), and in embarrassing social situations (Duncan & Tarulli, 2009). Despite reports that adults do, in fact, talk out loud to themselves, the possible beneficial effects of private speech in adults remain largely understudied, likely due to the original theory suggesting that the phenomenon disappears by adulthood, in addition to the fact that talking out loud to oneself has been associated with atypical development (Abdul Aziz et al., 2017; Mulvihill et al., 2023) and/or the folk psychology belief that it is a sign of mental illness or psychopathy (despite that claim lacking empirical support, see Glenn & Cunningham, 2000). Interestingly, this apparent under-appreciation regarding the benefits that private speech might confer on adult *cognitive* performance stands in contrast with there being substantial literature demonstrating the beneficial effects of private speech for *sports* performance, for example, when first learning to golf (Marshall et al, 2016; Turner et al., 2018, see Hatzigeorgiadis et al. 2011 for sports psychology review and meta-analysis, noting that some of the studies involved instructing learners to use *inner*, not private, speech). Similar benefits of private speech have been reported for *second language acquisition* (de Guerrero, 2018; Oxford, 1994).

Although there is a general dearth of studies investigating the relationship between private speech and cognitive performance in adults, there are two other kinds of literature that speak to the topic. The *first* is the "verbalization" literature, which shows that cognitive performance (e.g., working memory/executive function) is enhanced when participants are instructed to label objects out loud and/or name the task rule (see Schubert, 2022 and Souza & Skóra, 2017 for reviews in adults, and page 260 of Doebel & Zelazo, 2015 for a meta-analytic discussion of labeling/task naming effects in children)[2]. For example, Kray et al. (2008) investigated the benefits of verbalization on cognitive performance across the life span (young children = 7-9 years, older children = 11-13 years, young adults = 25-27 years, older adults = 66-77 years). In this study, they used a task-switching procedure, with performance represented by the reaction time difference between single and mixed blocks (referred to as the "mixing cost"). Using a within-subject design, performance was compared across conditions in which participants (a) named the next task to be performed (i.e. task-relevant verbalization), (b) verbalized words not related to the task at hand (i.e. task-irrelevant verbalization), or (c) did not verbalize (control condition, which can be considered the "baseline" condition). For all ages, mixing costs were substantially reduced under task-relevant verbalization and increased under task-irrelevant verbalization (compared to baseline). Interestingly, the benefit of task-relevant

[2] Interestingly, the improvements to working memory as a result of labeling out loud are opposite to another known effect, referred to as "overshadowing", in which describing an object out loud (for example, the bouquet of a wine) can hinder recognition memory for that object, especially if one possesses expertise in that domain (for example, a wine expert), see Chin and Schooler (2008) for review. The topic of overshadowing is outside the scope of this study, and will not be discussed further here.

speech was greatest for the two age groups (young children and older adults) whose baseline performance was the poorest, a finding that is relevant to the "Expertise Reversal Effect", discussed further below.

Although these previously-reported beneficial effects of verbalization bode well for there also being beneficial effects of private speech for cognitive performance in adults, it is important to point out that verbalization and private speech can differ along several dimensions and therefore may not be expected to show identical effects on cognitive performance. At the *phenomenological* level, private speech is a more natural and unrestricted process of "thinking out loud", and therefore is likely to be much richer (in both quantity and content) than simply labeling/naming out loud. At a *strategic* level, labeling/naming may be beneficial for simple tasks, while private speech may be beneficial for more complicated tasks, for example, ones that require spatial planning (like the Tower of London). Finally, at an *empirical* level, by not restricting the amount/content of self-talk, private speech studies are better positioned than verbalization studies to ask whether these quantitative/qualitative variables predict performance.

The *second* relevant literature comes from "articulatory suppression" studies, which show that suppressing (or at least diminishing) self-talk *impairs* performance on (some) cognitive tasks (see Fatzer & Roebers, 2012; Lidstone et al., 2010 for studies in children and Nedergaard et al., 2022 for a review in adults). In contrast to the verbalization literature (in children and adults) and private speech literature (in children), which suggest that talking out loud is a *sufficient* strategy for improving cognitive performance, the articulatory suppression literature suggests that self-talk may be a *necessary* element. Although articulatory suppression studies are relevant to the topic of private speech, it is important to point out that this paradigm is designed to suppress mainly *inner*, not private, speech. Like the case made above for different types of talking out loud (verbalization vs. private speech) being different on several dimensions, the same argument can be made when comparing inner vs. private speech. Specifically, the two speech types might differ phenomenologically (in their amount and/or content) and strategically (benefiting performance differentially depending on the task). Moreover, on an empirical level, only private speech can be measured objectively, thereby allowing a more rigorous investigation of its relationship with cognitive performance. Thus, while the results from articulatory suppression studies suggest an important role of inner speech in cognitive performance, much knowledge can be gained by studying the association between private speech and cognitive performance, about which little is known in adults.

To address this gap in the field, the main goal of the current study was to ask whether young adults' amount of private speech while performing a cognitive (visual-spatial working memory) task is positively associated with their performance on that task. The cognitive task consisted of a card-matching game, called "concentration cat" (iOS App), wherein players are tasked with finding hidden pairs of matching images within an array by tapping/revealing two cards at a time. For each participant, cognitive performance was measured in two "Private Speech" trials. In both, they were instructed to finish the game in as few turns as possible, while talking out loud to themselves as much as possible (without any restriction regarding the content). Unbeknownst to the participant, they were audio-recorded on these trials so that the amount (and content) of their private speech could later be determined. This design allowed us to investigate the within-person relationship between amount of private speech and performance, i.e., asking whether individuals performed better on trials for which they produced a greater amount of private speech. To our knowledge, this within-person approach has yet to be tested in

the adult private speech literature, although, as is the case for all correlational studies of private speech (see above), finding a positive correlation between amount of private speech and performance still leaves open the question of causality and the direction of causality, an issue we return to in the *Discussion*.

A secondary goal of the current study was to ask whether within-person associations between amount of private speech and performance (should they exist) vary depending on the *baseline competency* of the participant in the card-matching game. To obtain this measure, prior to the Private Speech trials, participants were asked to perform the same card-matching game under a condition where they were *not* instructed to talk out loud, which we refer to as the "Baseline" condition. Finding that Baseline performance *moderates* the relationship between amount of private speech and performance (on the Private Speech trials) would provide evidence for what is referred to as the Expertise Reversal Effect. This effect, which originated from educational psychology (Kalyuga, 2007), proposes that strategies for improving on a task may be beneficial for novices, yet less effective (or even harmful) for experts (as seen in Kray et al, 2008, mentioned above). A commonplace example is learning to tie one's shoes, which is a type of procedural memory. At first, using self-talk (with either inner or private speech) to explain the procedure ("make one loop, tie the other end around the loop, etc.") is helpful, but once one has become an expert in shoe-tying, then self-talk gets in the way. In fact, in the sports psychology literature (mentioned above), some studies report that talking out loud can hinder golf performance once people become experts (Beilock & Carr, 2001; Marshall et al., 2016). If a similar phenomenon exists for private speech, we expect that baseline competency on the task will moderate the relationship between private speech usage and performance.

Method

The hypothesis, study design, exclusion criteria, and analysis plan were preregistered: https://osf.io/jqfhc

Participants

Participants were undergraduate students recruited through a participant pool at UC San Diego, between February 2022 - September 2022. Eligibility was restricted to participants who reported being at least 18 years old. All participants gave their informed consent before participating and were compensated with course credit. The study was approved by the Institutional Review Board. The collected sample consisted of 120 participants. The sample size, which was determined by a priori power simulation, and exclusion criteria, are detailed in the pre-registration. Two participants were excluded. One was excluded because their performance in the Baseline condition was three standard deviations worse than the group average. The other was excluded because, at the end of the study, they did not consent to their audio recording being used for analysis. A total of 118 participants, ages 18 to 33 years ($M = 20.13$, $SD = 1.91$) were retained for analysis. Gender identities were 71.2% women, 26.3% men, and 2.5% non-binary. Ethnicities were 46.6% Asian, 19.5% White, 18.6% Hispanic, 4.2% Middle Eastern or North African, 2.5% Black/African American, 5.9% mixed, and 2.5% "prefer not to say".

Procedure

Card-Matching Task.

The study used a card-matching game, called “concentration cat” (iOS App), wherein players are tasked with finding hidden pairs of matching images within an array by tapping/revealing two cards at a time. If a match is made, those cards disappear. If instead there is a mismatch, those cards are automatically hidden again. This task relies on visual-spatial working memory, with the player needing to remember where in the array of cards they last saw an image. To play the game efficiently, the player aims to use as few “turns” as possible, with a turn defined as a pair of taps.

In the current study, we used the card-matching game in a 5 x 5 card array, which required 12 unique images, noting that each image is hidden under two cards, resulting in 24 total cards. Because a 5 x 5 array has 25 spots, one of those spots (i.e., the bottom/right spot of the array) was intentionally left empty. In the current study, each participant was tested on four trials, and thus we needed 48 unique images (i.e., 12 per trial). These were clip-art images, selected with the goal of having the images be easily labelable[3].

In-lab Procedure.

When a participant came to the lab, they entered a test room with an experimenter. To begin, they were told that the experiment involved playing a card-matching game, which was explained to them by having them watch a brief video demonstration of the game on a laptop computer. This “demo” video consisted of a 2 x 3 array of hidden cards, using images that were different from those used in the actual trials (below). The experimenter stopped the video now and then to elaborate on the rules and goals of the game. Then, the experimenter proceeded by setting up the participant to play four trials of the game on an iPad. The experimenter was outside the testing room during all four trials and only came back in between the trials to deliver instruction for the next trial, so as to not make the participant uncomfortable.

The first two trials were the “Baseline” condition, in which participants were asked to finish the game in as few turns as possible, noting that rarely ever did a participant spontaneously talk out loud in this condition (see *Results*). Performance on the two Baseline trials was averaged and used as a measure of *competency* on the task, to explore the “Expertise Reversal Effect” (see *Introduction*). Here, we assume that the variation observed in Baseline performance across our sample is a proxy for variations in expertise on the task. We refer to this variation as level of “competency”, rather than using the term “expertise”, since the latter is typically used to refer to the amount of *training* one has on a task, and this was not manipulated in our study.

In the next two trials, referred to as the "Private Speech” condition, participants were given the same instructions but were also asked to "talk out loud as much as possible" during the game. Specifically, they were instructed to:

> “Talk to yourself audibly or externally throughout the game or as much as you can. You can use whatever language you're comfortable with. We do not have instructions on the content of your self-talk. The volume of your self-talk can be comparable to the volume

[3] Because we had originally hoped to also test children, we wanted to make sure the images were labelable by children and adults. To this end, we selected words that are concrete nouns from the English dataset of WordBank, which is a database of children’s vocabulary development (Frank et al., 2017). Data were downloaded on February 19, 2020. We used nouns from Word Bank that can be produced by at least 65% of 30-month-olds, with the assumption that 100% of 4- to 6-year-olds (which was our original target age) would be able to produce these nouns. Once we determined the viable nouns, we then searched clip-art images of those nouns from Google. The clip-art patterns were all in color with white backgrounds.

of your social conversations. I (the experimenter) will be outside, and the door will be closed. I wouldn't be able to hear you during the game."

Unbeknownst to the participants, we recorded their speech output through an iPad microphone, so as to calculate an objective measurement of their amount and content of private speech (*see below*). Also unbeknownst to the participants, we used a screen capture function on the iPad to collect three pieces of information: (1) *number of turns*, and *(2) time* to complete the trial (automatically spit out by the iOS App after each trial) and (3) *sequence of card taps*. (1) was used as our main performance measure, (2) was used to compute *rate* of private speech, and (3) was used to compute a nuanced metric for performance (*see below*). The screen and audio recordings were collected for all four trials (the Baseline and Private Speech trials). At the end of the study, participants were debriefed about being secretly recorded during the experiment. They were given a consent form to indicate if they agreed for their audio to be analyzed for research purposes.

As part of our exploratory analyses, after each of the two Private Speech trials, we asked participants to answer "experiential" questions over Qualtrics on a laptop provided by the experimenter (e.g., comfort in talking out loud, self-reported amount of private speech), but these data are not presented in the body of this paper due to a lack of relevancy. A full list of experiential questions, and some exploratory analyses conducted on those questions (which were part of the pre-registration), are presented in Supplementary Material.

Measures

Performance Measurement.

The main measurement of performance for each trial was "*number of turns" (i.e., a pair of taps)* to finish the card-matching game. This measure is regarded as a straightforward and holistic evaluation of efficiency in the card-matching game (Krøjgaard et al., 2019), and is in line with many previous studies that used the same game (Eskritt & Lee, 2002; Washburn & Gulledge, 2002). However, because it has been suggested that it may be beneficial to use more nuanced performance metrics (see examples in Baker-Ward & Ornstein, 1988; Krøjgaard et al., 2019; and Schumann-Hengsteler, 1996), in the current study, in addition to using "number of turns", we used an additional metric that accounts for varying degrees of luck while playing the game (see Schmidt, 2005 for full details). This measure, which we refer to as the "performance ratio", divides the "number of turns" the participant uses to finish the game by the number of turns it would have taken assuming perfect memory (i.e., no memory errors, based on the tap choices of the participant). A ratio of 1.0 indicates perfect performance[4].

Amount of Private Speech (PS).

The audio recordings of participants' private speech were analyzed offline by the first author and her research assistants. For each of the two Private Speech trials, the audio recording was transcribed by the first author when the language was one she understood (English: 85.3% of trials, Mandarin: 10.5% of trials). On the occasion that participants spoke in a language other than those, we had research assistants or volunteers who spoke these other languages to help

[4] Note that "number of turns" was found to be highly correlated with the "performance ratio" ($r(599) = 0.884$, $p < 0.001$). The results of the current study are presented using "number of turns" (as this is what the field mostly uses), although brief mention of results using "performance ratio" are also presented.

transcribe (0.8% Arabic, 0.4% Burmese, 0.8% Korean, 0.8% Gujarati, 0.8% Spanish). Note that these percentages are out of the total number of trials, as some participants switched languages between their first and second Private Speech trials[5]. Data were entered into a spreadsheet in units of "Utterances", defined as an audible verbal unit that is separated by differences in semantic meaning *or* at least one second of temporal distance. For example, "Dog at the top right corner" would be considered one utterance, whereas "Is the dog here? Nope." would be considered two utterances (Frausel et al., 2020; Rowe, 2012; Rowe & Goldin-Meadow, 2009). Because, for some participants, we had a second transcriber (in addition to the first author), we were able to test inter-rater reliability. Data from 16 participants (32 trials) showed very high inter-rater reliability in quantifying amount of PS (ICC = 0.995).

In our previous pilot studies (see pre-registration), we calculated amount of PS in four different ways: 1) total number of words, 2) total number of utterances, 3) word rate (words/minutes), and 4) utterance rate (utterances/minutes) (with minutes calculated as the time to finish the task), and found that *utterance rate* was the best predictor of performance on the task. Thus, in the current study, we used utterance rate as our measure of amount of PS, noting that there are other reasons to use this particular measure. First, in the rare number of previous adult studies that measured amount of private speech (Duncan & Cheyne, 2001), they likewise employed utterance rate as their measure (and similarly, many teens/children studies use this measure, for instance, Fernyhough & Fradley, 2005; Kronk, 1994; Mulvihill et al. 2021). Second, in our exploratory analyses where we investigate the content of private speech (including categories such as "rehearsing" or "motivational", see *Results*), utterance is the only unit that makes sense. Finally, *utterance rate* is more appropriate than *total utterances*, as rate controls for variations in time to complete the task that might otherwise confound the results. For example, it is likely that poor performance will increase the time needed to finish the task, which in turn, is likely to result in more *total* utterances (especially when participants are explicitly instructed to talk out loud, as in the current study). This would then lead to the misleading conclusion that increasing amounts of private speech (in the form of *total* utterances) are associated with *poorer* performance[6]. Thus, it is more appropriate to use utterance rate, rather than total utterances.

In the *Results* section, we describe the model analyses performed with these variables, noting that all variables met our criterion of normality by passing a test of skewness (acceptable range -2 to 2) and kurtosis (acceptable range -2 to 2).

Data Transformation.

[5] Although we did not specifically ask participants about their primary language or other relevant questions for researchers interested in bilingualism, we did observe some language switching in our dataset. Specifically, we found that two participants switched languages - one Burmese speaker switched from Burmese to English, and one participant used a mix of English and Spanish during the first private speech trial, and only Spanish during the second private speech trial.

[6] These assumptions were, in fact, borne out in the data. Specifically, using multi-level modeling with one variable as an independent, and the other as a dependent, variable, we found that 1) the time to complete the task was negatively correlated with performance (i.e., the longer the time to complete the task, the worse the performance: $p < 0.001$), 2) the time to complete the task was positively correlated with total number of utterances (i.e., the longer the time to complete the task, the more total utterances that were made: $p < 0.001$), and 3) the total number of utterances was negatively correlated with performance (i.e., the more total utterances that were made, the worse the performance: $p = 0.003$).

Our use of two trials for the Private Speech condition allowed us to investigate *within*-person relationships between amount of PS and performance, i.e., asking whether an individual performed better on the trial for which they produced a greater amount of private speech. This is in contrast to analyzing the data using a *between*-person approach, i.e., asking whether performance was better for individuals who talked more vs. those who talked less. While both approaches (within- and between-person) are correlational in nature, and thus cannot prove causality, we chose the within-person approach because the between-person approach adds an additional challenge in discussions of causality; any observed between-subject correlation can be driven by a trait-based third variable, such as intelligence. That is, it could be that more intelligent people both talk out loud more and perform better. We return to the topic of causality, and future directions for testing causality, in the *Discussion*.

In order to conduct a within-person analysis within our multilevel models, we first person-mean centered the amount of PS. For example, if a participant's utterance/min was 40 on one trial and 20 on the other (with a mean of 30), this resulted in the amount of PS in their two Private Speech trials being encoded as +10 and -10, respectively. Note that in 0.9% of the Private Speech trials, the number of utterances was 0 (i.e., the participant did not follow the instructions to talk out loud), but values of 0 are permissible in the analyses. This person-centered transformation, sometimes referred to as "centering-within-cluster", reveals Level 1 (i.e., within-person) effects while eliminating Level 2 (i.e., between-person) effects in a multilevel model (Enders & Tofighi, 2007).

Finally, regarding the performance measures, both "number of turns" and "performance ratio" were z-transformed for easier comparison of effect sizes across different performance metrics (both within the current study and between the current and past/future studies).

Results

Descriptive analysis.

Of the 236 Private Speech trials (2 trials x 118 participants), three were excluded because the performance was three standard deviations worse than the trial-wise average performance for Private Speech trials. Note that this exclusion criterion was part of our pre-registration, and that missing data points of this sort are permissible in multilevel models (Huta, 2014). Of the remaining 233 Private Speech trials, 3.9% had a perfect memory performance (see Methods for definition). With regard to Baseline trials, which were used as a measure of baseline competency, 1.6% had perfect memory performance, and in only 0.4% of these trials did a participant make any spontaneous utterances[7]. Across the entire 233 useable Private Speech trials, the mean number of utterances/minute was 27.56 (*SD* = 11.26). In Table 1, we present mean utterances/mins, and mean performances, in terms of both "number of turns" and "performance ratios", separately for the first vs. second Private Speech trials. We separate the data by trial to show that there were no overall increases between the first and second trials (*p*-values for dependent t-tests for amount of PS and the two performance metrics were all > 0.72). This is important because it rules out the possibility that any relationship found between amount of PS and performance is a spurious result of an order effect (for example, which could happen if participants improved in their performance, *and* were more willing to talk out loud, between the first and second trial).

[7] Because it was a rare occurrence and the amount of utterance was quite low (on average, being in the 3rd percentile of that seen in the private speech condition), we did not exclude these trials.

Table 1

Means and standard deviations of Amount of Private Speech (utterance/minute), and the two ways to calculate performance: Number of Turns and Performance Ratio, separately for each of the two Private Speech trials

Variable	First Private Speech Trial		Second Private Speech Trial	
	M	*SD*	*M*	*SD*
Amount of Private Speech (utterance/min)[8]	27.31	10.54	27.80	12.08
Number of Turns (lower numbers = better performance)	25.29	4.11	25.10	3.98
Performance Ratio (lower numbers = better performance)	1.31	0.19	1.30	0.19

Testing the Relationship between Amount of Private Speech and Performance.

Using a Type III sum of squares multilevel regression model, we asked whether amount of PS predicts performance. In addition, we asked whether this relationship (if it exists) is stronger for those with poor baseline competency on the task, in line with the Expertise Reversal Effect (see *Introduction*). The dependent variable was performance (specifically, "number of turns" to complete the task) and the predictor terms were: 1) amount of PS (entered as a fixed effect), 2) baseline competency (entered as a fixed effect), and 3) the interaction between (1) and (2). For each participant, there were two Private Speech trials, and thus, the unit of analysis was "trial", with Participant included as a random intercept effect. Because we were interested in within-person effects, amount of PS was person-mean centered for each of the two Private Speech trials (see *Data Transformation in Methods).*

The results of this model, shown in Table 2, reveal three main findings. First, there was a main effect of amount of PS on performance (β = -0.04, 95% CI = [-0.06, -0.01], p = 0.003), with higher amounts of private speech being associated with fewer turns, i.e., better performance. Because this was a within-person analysis, this result means that participants performed better on the trial for which they produced a greater amount of private speech. Second, contrary to what one would expect from the Expertise Reversal Effect, the interaction between amount of PS and

[8] For comparison, the average amount of "instructed" private speech in our study (i.e., M = 27.50, SD = 11.36 utterances/minute) was substantially higher than the rate of "spontaneous" private speech reported in other studies of adults. For example, in Duncan & Cheyne (2001), M = 2.95, SD = 1.94 utterances/minute for their data entry tasks; M = 1.26, SD = 1.26 utterances/min for their paper-folding task. Further studies are needed to explore the differential effects (and content) of instructed vs. spontaneous private speech. For instance, spontaneous private speech may have more varying levels of internalization, compared with its prompted counterpart.

baseline competency was insignificant (p = 0.145), meaning that the relationship between amount of PS and performance was invariant across participants with different levels of baseline competency. Third, as might be expected, baseline competency predicted performance in the Private Speech condition, i.e., people who did better in the Baseline condition did better in the Private Speech condition (β = 0.43, 95% CI = [0.30, 0.56], p < 0.001). When we removed baseline competency from the model, the effects of amount of PS remained identical, although the marginal R-squared of the model necessarily became smaller (0.023, data not shown)[9].

Table 2

The results of a Type III Multilevel Model for Testing the Effects of Private Speech on Performance and an Expertise Reversal Effect

	Performance in the Private Speech Condition (number of turns)		
Predictors	Estimates	95% CI	*p*
(intercept)	-0.01	-0.14 – 0.12	0.886
Baseline Competency	0.43	0.30 – 0.56	<0.001
Amount of PS	-0.04	-0.06 – -0.01	0.003
Baseline Competency * Amount of PS	-0.02	-0.05 – 0.01	0.145
Random Effects			
σ2	0.54		
τ00	0.23 $_{\text{Participant}}$		
ICC	0.30		
N	117 $_{\text{Participant}}$		
Observations	228		
Marginal R^2 / Conditional R^2	0.219 / 0.453		

[9] When "performance ratio" was used as the performance metric, the results were nearly identical (as might be expected given that the two metrics – "number of turns" and "performance ratio" are highly correlated, see footnote 4). Specifically, there was a main effect of amount of PS (β = -0.03, 95% CI = [-0.06, -0.01], p = 0.017) and no significant interaction (p = 0.255).

Private Speech Content Distribution

As part of an exploratory analysis, we investigated the content of Private Speech, as such findings might steer future studies investigating the differential effects of different types of private speech. To this end, we placed each utterance into one of 14 categories, outlined in Table 3 (below). The categories were inspired by a mixture of those referenced in previous literature (Diaz, 2014; Duncan & Cheyne, 2001; Winsler, 2009), and additional categories we observed in our specific visual-spatial working memory task. Because, for some participants, we had a second transcriber (in addition to the first author), we were able to test inter-rater reliability. Data from 13 participants (26 trials) showed very high inter-rater reliability in quantifying *content* of private speech (κ = .808, parentage agreement [categorization of content being the same] = 90%, see Landis & Koch, 1977 for the use of Cohen's kappa [κ]). Next, for each trial, we calculated the frequency distribution of the different types of utterances observed in that trial. For instance, if a trial contained five labeling utterances, four negative emotional utterances, and one rehearsing/looping, this resulted in values of 50%, 40%, and 10%, for each one of those three categories, respectively, and values of 0% for the other 11 categories. In Figure 1, we plot the frequency distribution of the 14 utterance types across all trials. That is, each dot represents the frequency of a given utterance within a single trial. For each utterance category, we also show the mean and standard deviations of these values.

Table 3

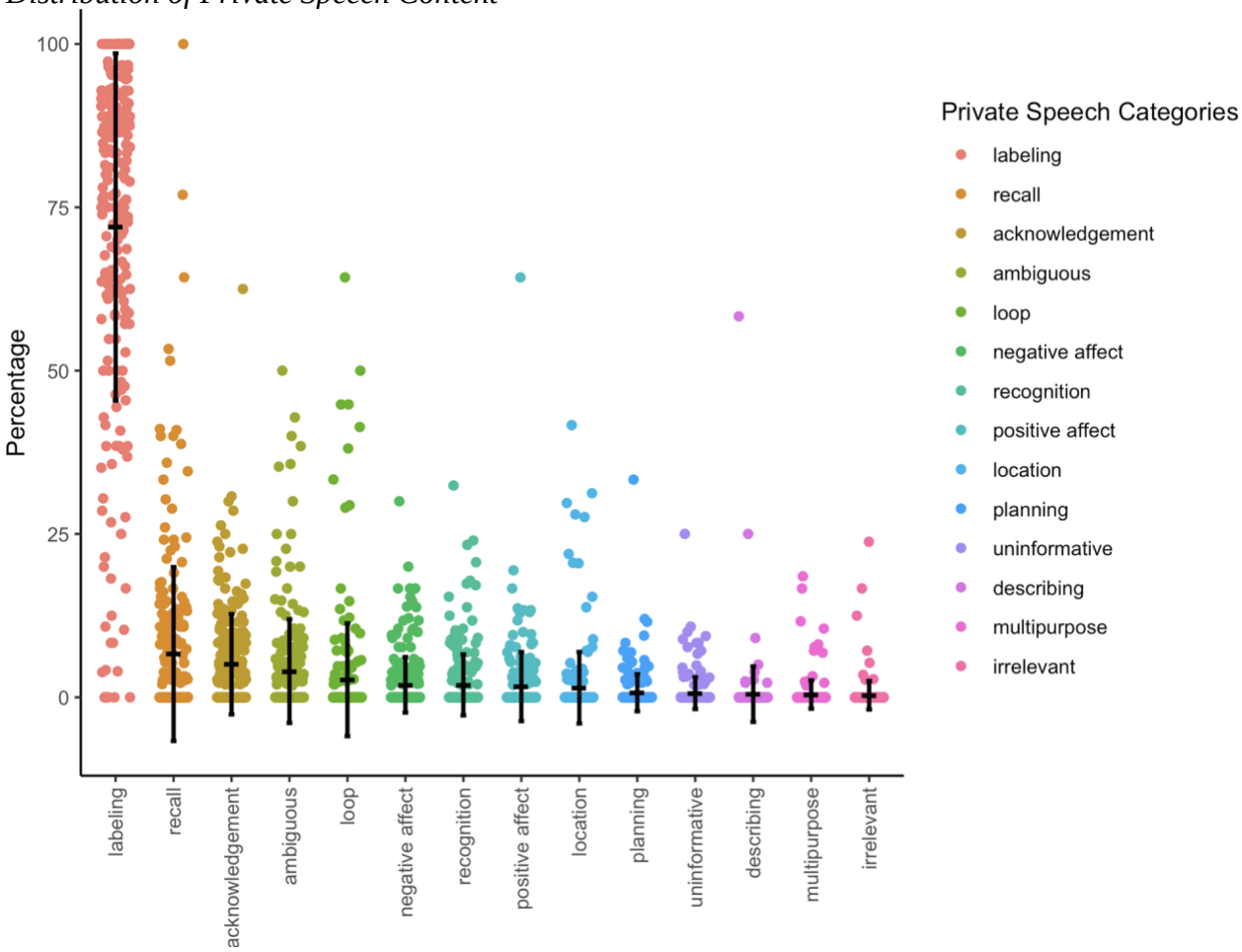
Private Speech Content Categories, Definitions, and Examples from the Current Dataset

Category	**Definition**	**Examples from the study**
Acknowledgment	Spontaneous reactions that are not emotional expressions.	"Ah", "ha", "I don't know what that is", "what?", "alright", and "ok"
Ambiguous or unclear	Audible but unintelligible whispering	The content cannot be coded, but the quantity was estimated.
Describing	Verbally describing stimuli, but no label	"A yellow and round… thing"
Irrelevant	Irrelevant to task completion.	"this is a kid thing", "that is cute"
Labeling	Labeling the card patterns or showing an attempt to label.	"Dog", and "apple"
Location	Including location terms, or directions	"Saw this one up here","the corner"
Multipurpose	Encoding both the location and card patterns aloud	"Elephant is in the middle", and "dog is top right"
Negative Affect Expression	Expressing pessimism, discouragement, and criticism.	"Looks like I messed up already", "oh man!", and profanities.
Planning	Planning for actions. Self-guided, self-managing attempts.	"Ok start from the top right", "do not tap this", "going to try the edges"
Positive Affect	Expressing optimism, encouragement,	"Good job", and "that's getting

Expression	and praise	better"
Recall	After seeing an old card, trying to recall where they saw the card last time or failing to recall.	“Where is the button, I do not know", "I saw a cow somewhere over here”
Recognition	Trying to recognize or to figure out if they have seen this card before. Recognition is assumed to take place before recall.	"Just saw that one", "don't think I've clicked this one yet", and "this isn't tapped"
Rehearsing	Rehearsing the previously seen stimuli when revealing new cards.	"Cat, bathtub, key, dog, blanket”
Uninformative	Not serving any specific function other than showing the individual is paying attention to the game.	"dudududu", "this one", and "let's see"
Irrelevant	Irrelevant to task completion.	"this is a kid thing", "that is cute”
Other	Utterances that do not belong to any of the categories above.	

Figure 1

Distribution of Private Speech Content



Note. Distribution of Private Speech Content. Note. The categories are ordered along the X-axis from the highest frequency (leftmost) to the lowest frequency (rightmost). The horizontal bars are the means, and the ranges of the vertical lines are the standard deviations of the private speech content categories.

The results of this analysis revealed that "labeling" was the most frequent category (for example, "dog", "house"), with a mean frequency of 71.9%. As we argue in the *Discussion*, the phenomenon of labeling is likely to be a type of strategizing to remember the location of the matching pair (as opposed to being a *response* to making a correct/incorrect match). In a similar vein, many of the other utterance types (for example, "recall", which consisted of phrases like "I saw a cow somewhere over here") seemed to be strategic in nature, that is, occurring *prior* to a correct/incorrect match. The mean frequency across all categories that appear to be strategic (including "labeling", "planning", "recall", "recognition", "looping", and "describing") was 84.39%. In contrast to strategizing, categories that involved "positive affect" (for example, "good job") and "negative affect" (for example, "looks like I messed up") seemed to occur in response to (good or poor) performance. The mean frequency of these affective response categories was

3.57%, with roughly half being positive, and half being negative, responses. Note that 12.04% of utterances fell neither into strategizing nor responding.

Discussion

The results of the current study conducted in young adults show that the degree to which one uses private speech, when instructed to do so, is positively associated with performance on a cognitive task, specifically, a visual-spatial working memory task. In addition, the strength of this relationship is not moderated by baseline competency on the task. Before moving on with a discussion of how private speech might benefit performance, we must address the fact that the results of the current study are correlational in nature, and therefore present some challenges in establishing a causal link. The current study tested participants on two Private Speech trials, and then, using multilevel modeling, asked whether participants performed better on Private Speech trials in which they talked out loud more. We chose this within-subject analysis because it is less prone to "third variable" explanations associated with using a between-person approach. For example, in a between-subjects analysis, if participants who talk out loud more also perform better, this association could be driven by a trait-based variable, such as intelligence. While such a *trait*-based explanation is removed in a within-subject design, there is still the possibility of a *state*-based third variable, such as *compliance*, underlying the association. For example, if participants in our study tried harder to follow directions on the second of the two Private Speech trials, and their level of compliance was mirrored on *both* tasks (Task 1 = perform the card-matching game as efficiently as possible *and* Task 2 = talk out loud as much as possible), this could underlie the observed relationship between amount of PS and performance. At least at a *group* level, this does not appear to be the case, as we found no order effects between the first and second Private Speech trials (see *Results*). Of course, it is still possible that such order effects exist at the *individual* level, yet in opposite directions across participants (thus canceling out at the group level). If this were the case, then there still exists the possibility of a state-based third variable (like compliance) underlying the observed relationship between amount of PS and performance, without there being a causal relationship between the two.

Finally, even if the correlational relationship between amount of PS and performance *is* the result of a causal relationship between the two, the *direction* of causality is uncertain; talking out loud more might lead to improved performance, or conversely, people may talk out loud more in response to performing well. We believe that, in the current study, the former is more likely based on the *content* of participants' utterances while playing the game. As reported in *Results*, the vast majority of utterances (84.39%) appeared to be strategic in nature, in some way helping participants to remember the location of the matching pair. From this, we assume that the vast majority of utterances occurred *prior* to a correct/incorrect match. By contrast, a very small fraction (3.57%) of utterances appeared to be *affective responses* to a correct/incorrect match[10]. In sum, based on the content of participants' private speech, we think the most likely direction of causality – given that there *is* a causal relationship -- is that increased private speech led to improved performance. With respect to what might underlie the beneficial effects of private

[10] However, even if the proportion of "affective response" utterances had been substantial (which was not the case), the fact that half of these utterances were in response to good, and half in response to poor, performance, would end up cancelling each other out when looking at the relationship between amount of PS and performance. Specifically, performance would be positively associated with amount of *positive affect* private speech, yet negatively associated with amount of *negative affect* private speech.

speech on performance, we propose two potential mechanisms. First, as discussed in the *verbalization* literature, the act of labeling out loud (which was the most frequent type of utterance in the current study) may enhance working memory for objects through the activation of long-term categorical representations (see Souza & Skóra, 2017 for review). Second, as discussed in the *sports psychology* literature, the use of private speech may serve to increase attention to the task at hand, thus enhancing performance (see Hatzigeorgiadis & Galanis, 2017).

Still, because the current study was correlational in nature, the results cannot provide conclusive evidence that private speech benefits performance. As outlined in the *Introduction*, the obvious way to establish causality is to employ an *experimental* approach, comparing performance between conditions where participants are *instructed* vs. *not instructed* (or explicitly told not) to talk out loud. However, if one is to use this approach, careful consideration must be placed on how best to counterbalance conditions. Despite the fact that the current study measured performance in the two conditions required for an experimental approach (i.e., the Private Speech, and the Baseline, condition), it was not set up to *compare* the two since their order was not counterbalanced across participants. In designing our study (see pre-registration), the Baseline condition was included as a way to obtain a *trait* measure of competency on the task, so that we could determine whether it moderated the relationship between amount of PS and performance (discussed further, *below*). We tested the Baseline condition first because we were concerned that, if we randomized the order of the two conditions, participants who were tested in the Private Speech condition in the first block might feel they ought to talk out loud in the (subsequent) Baseline condition, which we did not want (see Turner et al. 2018, above, for similar logic in studies of sports performance). Based on the design of our study, comparisons between our Baseline and Private Speech conditions may be confounded by order effects, which could be in the form of a "fatigue effect" (a tendency to perform *worse* in the second condition) or a "practice effect" (a tendency to perform *better* in the second condition). Given that there is a true benefit of private speech on performance, a "fatigue effect" will result in an underestimate, and a "practice effect" will result in an overestimation, of this beneficial effect[11]. For this reason, the current study did not plan a comparison analysis between the Private Speech and Baseline conditions, however, future studies should plan to do so.

As noted above, we included the Baseline condition so that we could ask whether the relationship between amount of PS and performance was stronger for those with poor baseline competency on the task. This "Expertise Reversal Effect" proposes that strategies for improving on a task may be beneficial for novices, yet ineffective (or even harmful) for experts, on that task (see *Introduction*). However, we suggest that this phenomenon should be expanded to refer to the *relationship* between participant expertise and task difficulty, noting that either dimension can be manipulated within a study. For example, some studies investigate the benefits of talking out loud on performance by testing individuals with different levels of expertise on the *same* task (e.g., testing people of different ages, with the assumption that adults are more expert/competent than children, as in Kray et al., 2008, see *Introduction*), while other studies vary task difficulty amongst individuals presumed to have the *same* expertise (Fernyhough & Fradley, 2005). As such, when investigating the effects of private speech on cognitive performance, the following prediction can be made; private speech will help if the task is relatively hard for a given

[11] Although we did not find any systematic order effect between the two Private Speech trials (see Table 1 of Results), an order effect could nonetheless exist between the first two (Baseline) trials and the next two (Private Speech) trials.

individual, and not help (or even hurt) if the task is relatively easy for an individual. In addition, one should consider the fact that merely instructing participants to talk out loud while performing a cognitive task might be experienced as difficult because of the "dual-task" nature of the situation. That is, for those who find it difficult and/or uncomfortable to talk out loud, the increased cognitive load of the dual-task might negatively affect cognitive performance (see Jackson et al., 2023, Rhodes et al., 2019 for evidence that dual-tasks impair memory performance).

In the current study, where task difficulty was kept constant, we assume that variations in Baseline performance on the task reflect variations in how difficult the task was across participants, which we refer to as "competency". If this assumption is correct, our finding of no moderating effect of "competency" on the positive relationship between amount of PS and performance might be explained by positing that most participants were at a "sweet spot" regarding the relationship between task difficulty and their competency. Alternatively, it could be that there was a non-linear (inverted U-shaped) effect of competency on the relationship between amount of PS and performance, which we missed by using linear models (see Fernyhough & Fradley, 2005 for an inverted U-shape function between task difficulty and amount of PS, although they did not find an inverted U-shape function between task difficulty and the benefit of private speech). Future studies that systematically vary the relationship between task difficulty and expertise/competence (and perhaps use non-linear interaction terms, see Karaca-Mandic et al., 2012) will be required to address these possibilities.

On a final note, future studies should consider other variables that might affect the degree to which a person is benefited by using private speech. Task difficulty is an obvious variable to investigate, noting that the difficulty of the current card-matching game can easily be manipulated by changing the number of cards and/or the degree to which the images on those cards are labelable. Looking at the *content* of private speech in these different scenarios might shed light on underlying mechanisms of beneficial effects, as we know from previous studies that the content of spontaneous private speech varies with the nature of a task, and, in a reciprocal fashion, that instructing different types of verbalization (task-relevant vs. task-irrelevant) differentially affects performance (see *Introduction*).

Another variable of interest is one's *comfort level* in talking out loud, particularly when one is instructed to do so, as in the current study. It had originally been our intention (see pre-registration) to include comfort level in talking out loud as a potential moderator of the relationship between amount of PS and performance. As described in Supplementary materials, our method for determining comfort level was to ask participants, after each of the two Private Speech trials, to report (on a Likert scale) how comfortable they were talking out loud on that trial. Our hope was that we could use this experiential question as a *trait* measure of comfort in talking out loud (akin to how we used Baseline performance as a trait measure of *competency*). However, we ended up not including comfort level in the current analysis because it seemed unreliable; there was a fairly low correlation ($r = 0.30$), between participants' comfort responses on their first vs. second Private Speech trial. One explanation for this low reliability is that participants' reports of comfort level could have been confounded by how well they felt they performed on the card-matching task, as opposed to being a pure reflection of their comfort level in talking out loud. For example, after struggling to find the hidden pairs on a given trial, and then being asked about their comfort level in talking out loud, a participant may have inadvertently reported discomfort that was tied more to their performance than to their talking

out loud. For this reason, future studies investigating the effects of comfort in talking out loud should use an established trait-level measure like the Self-Talk Scale developed by Brinthaupt et al. (2009).

Lastly, the effect of *age* is another variable that can be investigated. The card-matching game of the current study was deliberately chosen because it can easily be administered in children (Krøjgaard et al., 2019), noting that we were careful to select images that we knew could be labeled by young children (see footnote 3). As such, future studies might map out the developmental trajectory - from young children to aging adults, of the effects observed in the current study. Determining the "when and how" private speech benefits cognitive performance (in all ages) may have important implications for real-world educational/instructional settings, a notion that has already been adopted for those learning a new sport or a second language.

References

Abdul Aziz, S., Fletcher, J., & Bayliss, D. M. (2017). Self-regulatory speech during planning and problem-solving in children with SLI and their typically developing peers. *International Journal of Language & Communication Disorders*, *52*(3), 311–322. https://doi.org/10.1111/1460-6984.12273

Alarcón-Rubio, D., Sánchez-Medina, J. A., & Winsler, A. (2013). Private speech in illiterate adults: Cognitive functions, task difficulty, and literacy. Journal of Adult Development, 20, 100-111.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. Psychological Bulletin, 141(5), 931–965. https://doi.org/10.1037/bul0000021

Al-Namlah, A. S., Meins, E., & Fernyhough, C. (2012). Self-regulatory private speech relates to children's recall and organization of autobiographical memories. *Early Childhood Research Quarterly*, *27*(3), 441–446. https://doi.org/10.1016/j.ecresq.2012.02.005

Baker-Ward, L., & Ornstein, P. A. (1988). Age differences in visual-spatial memory performance: Do children really out-perform adults when playing Concentration? *Bulletin of the Psychonomic Society*, *26*(4), 331–332. https://doi.org/10.3758/BF03337672

Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General*, *130*(4), 701–725. https://doi.org/10.1037/0096-3445.130.4.701

Berk, L. E. (1986). Relationship of elementary school children's private speech to behavioral accompaniment to task, attention, and task performance. *Developmental Psychology*, *22*(5), 671–680. https://doi.org/10.1037/0012-1649.22.5.671

Brinthaupt, T. M., Hein, M. B., & Kramer, T. E. (2009). The self-talk scale: Development, factor analysis, and validation. *Journal of Personality Assessment, 91*(1), 82-92.

Chin, J. M., & Schooler, J. W. (2008). Why do words hurt? Content, process, and criterion shift accounts of verbal overshadowing. *European Journal of Cognitive Psychology*, *20*(3), 396–413. https://doi.org/10.1080/09541440701728623

de Guerrero, M. C. M. (2018). Going covert: Inner and private speech in language learning. *Language Teaching*, *51*(1), 1–35. https://doi.org/10.1017/S0261444817000295

Diaz, R. (2014). Methodological Concerns in the Study of Private Speech. In R. M. Diaz (Ed.), *Private Speech* (pp. 65–92). Psychology Press. https://doi.org/10.4324/9781315807270-5

Doebel, S., & Zelazo, P. D. (2015). A meta-analysis of the Dimensional Change Card Sort: Implications for developmental theories and the measurement of executive function in children. Developmental Review, 38, 241-268.

Duncan, R. M., & Cheyne, J. A. (2001). Private speech in young adults: Task difficulty, self-regulation, and psychological predication. *Cognitive Development*, *16*, 889–906. https://doi.org/10.1016/S0885-2014(01)00069-7

Duncan, R., & Tarulli, D. (2009). On the persistence of private speech: Empirical and theoretical considerations. In A. Winsler, C. Fernyhough, & I Montero (Eds.), *Private speech, executive functioning, and the development of verbal self-regulation* (pp. 176–187). Cambridge University Press.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological methods*, *12*(2), 121.

Eskritt, M., & Lee, K. (2002). “Remember Where You Last Saw That Card”: Children’s Production of External Symbols as a Memory Aid. *Developmental Psychology, 38*, 254–266. https://doi.org/10.1037//0012-1649.38.2.254

Fatzer, S. T., & Roebers, C. M. (2012). Language and Executive Functions: The Effect of Articulatory Suppression on Executive Functioning in Children. *Journal of Cognition and Development*, *13*(4), 454–472. https://doi.org/10.1080/15248372.2011.608322

Fernyhough, C. (2004). Alien voices and inner dialogue: Towards a developmental account of auditory verbal hallucinations. *New Ideas in Psychology*, *22*(1), 49–68. https://doi.org/10.1016/j.newideapsych.2004.09.001

Fernyhough, C., & Fradley, E. (2005). Private speech on an executive task: Relations with task difficulty and task performance. Cognitive Development, 20(1), 103–120. https://doi.org/10.1016/j.cogdev.2004.11.002

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data*. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Frauenglass, M. H., & Diaz, R. M. (1985). Self-regulatory functions of children’s private speech: A critical analysis of recent challenges to Vygotsky’s theory. *Developmental Psychology*, *21*, 357–364. https://doi.org/10.1037/0012-1649.21.2.357

Frausel, R. R., Silvey, C., Freeman, C., Dowling, N., Richland, L. E., Levine, S. C., Raudenbush, S., & Goldin-Meadow, S. (2020). The Origins of Higher-Order Thinking Lie in Children’s Spontaneous Talk Across the Pre-School Years. *Cognition*, *200*, 104274. https://doi.org/10.1016/j.cognition.2020.104274

Glenn, S. M., & Cunningham, C. C. (2000). Parents' reports of young people with Down syndrome talking out loud to themselves. Mental Retardation, 38(6), 498-505.

Hatzigeorgiadis, A., & Galanis, E. (2017). Self-talk effectiveness and attention. *Current Opinion in Psychology, 16*, 138–142. https://doi.org/10.1016/j.copsyc.2017.05.014

Hatzigeorgiadis, A., Zourbanos, N., Galanis, E., & Theodorakis, Y. (2011). Self-Talk and Sports Performance: A Meta-Analysis. *Perspectives on Psychological Science*, *6*(4), 348–356. https://doi.org/10.1177/1745691611413136

Huta, V. (2014). When to use hierarchical linear modeling. The quantitative methods for psychology, 10(1), 13-28

Jackson, K. M., Shaw, T. H., & Helton, W. S. (2023). The effects of dual-task interference on visual search and verbal memory. Ergonomics, 66(1), 125-135.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational psychology review, 19, 509-539.

Karaca-Mandic, P., Norton, E. C., & Dowd, B. (2012). Interaction terms in nonlinear models. Health services research, 47(1pt1), 255-274.

Kray, J., Eber, J., & Karbach, J. (2008). Verbal self-instructions in task switching: a compensatory tool for action-control deficits in childhood and old age? Developmental Science, 11(2), 223-236.

Krøjgaard, P., Sonne, T., Lerebourg, M., Lambek, R., & Kingo, O. S. (2019). Eight-year-olds, but not six-year-olds, perform just as well as adults when playing Concentration: Resolving the enigma? *Consciousness and Cognition*, *69*, 81–94. https://doi.org/10.1016/j.concog.2019.01.015

Kronk, C. M. (1994). Private speech in adolescents. Adolescence, 29(116), 781.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159. https://doi.org/10.2307/2529310

Lee, J. (1998). The effects of five-year-old preschoolers' use of private speech on performance and attention for two kinds of problem-solving tasks (Order No. 9932671). Available from ProQuest Dissertations & Theses A&I; ProQuest Dissertations & Theses Global. (304413779). https://www.proquest.com/dissertations-theses/effects-five-year-old-preschoolers-use-private/docview/304413779/se-2

Lidstone, J. S. M., Meins, E., & Fernyhough, C. (2010). The roles of private speech and inner speech in planning during middle childhood: Evidence from a dual task paradigm. *Journal of Experimental Child Psychology*, *107*(4), 438–451. https://doi.org/10.1016/j.jecp.2010.06.002

Marshall, D. V., Hanrahan, S. J., & Comoutos, N. (2016). The Effects of Self-Talk Cues on the Putting Performance of Golfers Susceptible to Detrimental Putting Performances Under High Pressure Settings. *International Journal of Golf Science*, *5*(2), 116–134.

Müller, U., Zelazo, P. D., Hood, S., Leone, T., & Rohrer, L. (2004). Interference control in a new rule use task: Age-related changes, labeling, and attention. Child development, *75*(5), 1594-1609.

Mulvihill, A., Matthews, N., Dux, P. E., & Carroll, A. (2021). Preschool children's private speech content and performance on executive functioning and problem-solving tasks. *Cognitive Development*, *60*, 101116. https://doi.org/10.1016/j.cogdev.2021.101116

Mulvihill, A., Matthews, N., & CARROLL, A. (2023). Task difficulty and private speech in typically developing and at-risk preschool children. Journal of Child Language, 50(2), 464-491.

Nedergaard, J. S., Wallentin, M., & Lupyan, G. (2023). Verbal interference paradigms: A systematic review investigating the role of language in cognition. Psychonomic Bulletin & Review, 30(2), 464-488.

Oxford, R. (1994). *Language Learning Strategies: An Update. ERIC Digest*. ERIC/CLL, 1118 22nd Street, N. https://eric.ed.gov/?id=ED376707

Rhodes, S., Jaroslawska, A. J., Doherty, J. M., Belletier, C., Naveh-Benjamin, M., Cowan, N., ... & Logie, R. H. (2019). Storage and processing in working memory: Assessing dual-task performance and task prioritization across the adult lifespan. Journal of Experimental Psychology: General, 148(7), 1204.

Rowe, M. L. (2012). A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development*, *83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Rowe, M. L., & Goldin-Meadow, S. (2009). Early gesture selectively predicts later language learning. *Developmental Science*, *12*(1), 182–187. https://doi.org/10.1111/j.1467-7687.2008.00764.x

Sawyer, J. (2017). I think I can: Preschoolers' private speech and motivation in playful versus non-playful contexts. Early Childhood Research Quarterly, 38, 84-96.

Schmidt, A. (2005). *Remembering the concentration game: Chance or memory?* [Master of Arts, San Jose State University]. https://doi.org/10.31979/etd.pjgz-v6w4

Schubert, T. (2022). Labels aid visual working memory. *Nature Reviews Psychology*, *1*(12), Article 12. https://doi.org/10.1038/s44159-022-00135-y

Schumann-Hengsteler, R. (1996). Children's and Adults' Visuospatial Memory: The Game

Concentration. *The Journal of Genetic Psychology*, *157*(1), 77–92. https://doi.org/10.1080/00221325.1996.9914847

Soskin, W. F., & John, V. P. (1963). The Study of Spontaneous Talk. In R. G. Barker (Ed.), *The stream of behavior: Explorations of its structure & content* (pp. 228–281). Appleton-Century-Crofts. https://doi.org/10.1037/11177-012

Souza, A. S., & Skóra, Z. (2017). The interplay of language and visual perception in working memory. *Cognition*, *166*, 277–297. https://doi.org/10.1016/j.cognition.2017.05.038

Turner, M. J., Kirkham, L., & Wood, A. G. (2018). Teeing up for success: The effects of rational and irrational self-talk on the putting performance of amateur golfers. *Psychology of Sport and Exercise*, *38*, 148–153. https://doi.org/10.1016/j.psychsport.2018.06.012

Vygotsky, L. (1987). Thinking and Speech. *The Collected Works of L. S. Vygotsky*, *1*, 39–285.

Washburn, D. A., & Gulledge, J. P. (2002). A Species Difference in Visuospatial Memory in Adult Humans and Rhesus Monkeys: The Concentration Game. *International Journal of Comparative Psychology*, *15*(4). https://doi.org/10.46867/C43W25

Winsler, A. (2009). Still talking to ourselves after all these years: A review of current research on private speech. In A. Winsler, C. Fernyhough, & I Montero (Eds.), *Private speech, executive functioning, and the development of verbal self-regulation* (pp. 3–41). Cambridge University Press. https://doi.org/10.1017/CBO9780511581533.003

Winsler, A., De León, J. R., Wallace, B. A., Carlton, M. P., & Willson-Quayle, A. (2003). Private speech in preschool children: Developmental stability and change, across-task consistency, and relations with classroom behaviour. *Journal of Child Language*, *30*(3), 583–608. https://doi.org/10.1017/S0305000903005671

Winsler, A., Manfra, L., & Diaz, R. M. (2007). “Should I let them talk?”: Private speech and task performance among preschool children with and without behavior problems. Early Childhood Research Quarterly, *22*(2), 215-231.

Supplementary Material

Content:

1. **Description of All Experiential Variables**
2. **Descriptive Statistics of All Experiential Variables**
3. **Results on the exploratory question: Are subjective measures of amount of PS (based on participant self-report) a good substitute for objective measures of amount of PS (based on audio recordings)?**

1. Description of all Experiential Variables

Below is a full list of experiential variables that were preregistered and collected, but they are not presented in this paper due to a lack of relevancy.

(1) Extent: a self-estimation of the extent of their private speech usage (scale of 1 - 7, with 1 labeled as "Not at all" and 7 labeled as "Completely/Entirely"); (2) Extent Confidence: their level of confidence about the estimation in (1) (scale of 1 - 7, with 1 labeled as "Not at all" and 7 labeled as "Completely/Entirely"); (3) Percentage: a self-estimation of their private speech usage in a percentage (scale of 0 - 100%); (4) Percentage Confidence: their level of confidence about the estimation in (3) (scale of 1 - 7, with 1 labeled as "Not at all" and 7 labeled as "Completely/Entirely"); (5) Comfort Level: their comfort level of following the instruction to talk to oneself out-loud during the trial (scale of 1 - 7, with 1 labeled as "Completely uncomfortable" and 7 labeled as "Completely comfortable"); (6) Labeling: the extent to which their private speech during the trial was about labeling the card patterns (scale of 1 - 7, with 1 labeled as "Not at all" and 7 labeled as "Completely/Entirely"); (7) Positive Affect: the extent to which their private speech during the trial was about expressing positive affect (scale of 1 - 7, with 1 labeled as "Not at all" and 7 labeled as "Completely/Entirely"); (8) Negative Affect: the extent to which their private speech during the trial was about expressing negative affect (scale of 1 - 7, with 1 labeled as "Not at all" and 7 labeled as "Completely/Entirely"); (9) Language: the language they used when speaking out loud. After answering these online questions, the experimenter asked the participants two open-ended questions and took notes on the same spreadsheet that recorded their performance. The first question was "Did you use any strategy during the game? It is ok if you did not use any." The second question was "Did you notice any trend or change in your strategy across the four trials?". Some, but not all, of these variables, which were preregistered for exploratory analyses, are presented in this paper.

Note that (1) - (5) were experiential questions about the specific private speech trials and were asked twice for each participant: once immediately after each of the two private speech trials. Whereas (6) - (8) were about the overall experience of the private speech trials and were asked once or after the last private speech trial. The rest of the questions were open-ended questions and were not coded qualitatively and are not reported here. Rather, they were purely exploratory and were used to give the researchers a better understanding of participants' experiences to inform future private speech studies.

2. Descriptive Statistics of All Experiential Variables

Supplementary Table 1 and Supplementary Table 2 show the descriptive statistics of experiential questions related to specific private speech trials and variables that assess participants' overall experience, respectively.

Supplementary Table 1

Means and Standard Devisions of the Experiential Questions about Each of the Two Private Speech (PS) Trials

Variables	M (SD) of the 1st PS Trial	M (SD) of the 2nd PS Trial
Extent of PS Usage	5.01 (1.34)	5.69 (1.24)
Confidence with Their Own Extent of PS Usage Rating (above)	6.08 (0.97)	6.27 (1.02)
Percentage of time PS was used	71.86 (22.14)	77.65 (21.77)
Confidence with Their Own Percentage of PS Rating (above)	5.97 (0.98)	6.05 (0.97)
Comfort Level with PS during the Trial	4.62 (1.72)	5.15 (1.77)

Supplementary Table 2

Mean and Standard Deviations of the Experiential Questions Asked After the Last Trial

Variables	M (SD)
Extent of Labeling in PS	6.18 (1.35)
Extent of Positive Affect Expression in PS	2.35 (1.70)
Extent of Negative Affect Expression in PS	1.93 (1.26)

Note. The ratings of the questions in this table are all on 7-point Likert scales. The content distribution reported in the main manuscript was data from audio recordings.

3. *Are subjective measures of* amount of PS *(based on participant self-report) a good substitute for objective measures of* amount of PS *(based on audio recordings)*?

Here, we asked whether our subjective measure of the amount of private speech might be a good substitute for the objective measure obtained with audio recordings. Winsler & Nagleiri (2003) tested the association between 5-to-7-years olds awareness of their (spontaneous) private speech (Yes vs. No) and observed private speech (Yes vs. No), and found a significant phi correlation between the two. This means that even children are aware of their audible spontaneous self-talk. Therefore, we expect a significant positive association between self-reported and observed private speech in the sample of young adults.

Subjective amount of PS was measured with two questions right after a Private Speech trial, both of which started with “We realize we asked you to talk to yourself out loud as much as

you can during the game, but still, people differ in how much they do this". In the "extent" question, this was followed with "With this in mind, please let us know…..during the game, *how much* of the time were you talking out loud to yourself?' on a 7-point scale with 1 labeled as "Not at all" and 7 labeled as "completely/entirely". In the "percentage" question, this was followed with "With this in mind, please let us know ….during the game, what *percentage* of the time were you talking out loud to yourself?", with 0% and 100% being the endpoints.

As a first step, we asked whether the two types of subjective measures ("Extent" and "Percentage") were associated with each other, by using the same mixed-effect models (above) and asking how well "percentage" predicts "extent". Because the two were found to be significantly and strongly associated ($\beta = 0.82$, $p < 0.001$, 95% CI = [0.74, 0.90]), this suggests that the subjective measure is quite reliable. For this reason, all subsequent analyses were performed using just one of the two subjective measures, specifically, "Extent". *Next*, we asked whether the subjective and objective measures of amount of PS were associated with each other, by using the same mixed-effect models (above) and asking how well the "subjective" measure (entered as a predictor variable) predicts the "objective" measure (entered as the dependent variable).

Here, we asked how well the objective measure of amount of PS was correlated with the subjective measure, noting that only the objective measure was used in our models (see *Methods*). The results of a linear mixed-effect model revealed a significant association between objective and subjective measures ($\beta = 0.03$, $p = 0.040$, 95%CI = [0.00, 0.06], see Supplementary Table 3). While the association is significant, the beta is weak enough to suggest that subjective measures are *not* a good substitute for objective measures. One possibility is that this weak association results from low reliability in one or both of the measures. We think this is an unlikely explanation, however, since the "Extent vs. Percentage" analysis suggests good reliability for the subjective measure, and inter-rater tests suggest good reliability in the coding of the objective data (see *Methods*, above). More likely, subjective and objective measures are tapping into two different constructs. For example, in the subjective measure, participants may be reporting how much they feel they talked out loud *relative* to their own personal benchmark, which may or may not align with the objective truth. In sum, one might use caution when deciding whether or not to substitute objective with subjective measures (see *Discussion*).

Supplementary Table 3
Association between (Level 1) Objective and Subjective Extent of Private Speech.

	Extent of PS Usage		
Predictors	*Estimates*	*CI*	*p*
(intercept)	5.39	5.38-5.60	< 0.001
Objective amount of PS	0.03	0.00 – 0.06	0.040
Random Effects			

σ^2	0.83
$\tau_{00\ Participant}$	0.89
ICC	0.52
N	117
Observations	229
Marginal R^2 / Conditional R^2	0.009 / 0.520

Note. The objective amount of PS is the centered-within-cluster amount of utterance per minute.