UNIVERSITY OF CALIFORNIA SAN DIEGO

Speaking to Ourselves: Establishing the Cognitive Benefit of Private Speech in Young Adults

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Experimental Psychology

by

Xinqi Guo

Committee in charge:

Professor Karen Dobkins, Chair
Professor David Barner
Professor Benjamin Bergen
Professor Gail Heyman
Professor John Wixted

2023

The Dissertation of Xinqi Guo is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

I would like to dedicate this dissertation to my parents, who prioritize my happiness more than their own.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PS	Private Speech
IS	Inner Speech

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Karen Dobkins. Over the past six years, her unwavering support and guidance have been instrumental in my academic journey. The more time I spent in grad school, the more I realized the significance of having an advisor like Karen who genuinely cares about her students' well-being. Further, her approach to writing transformed my view of it from a daunting task to an engaging puzzle.

My appreciation extends to my dissertation committee members: Dr. David Barner, Dr. Benjamin Bergen, Dr. Gail Heyman, and Dr. John Wixted. Their expertise, flexibility, and patience have been highly invaluable.

I'm grateful for the constant support from my lab mates: Debra Lindsay, Stefanie Holden, Stephen Raynes, Silvia Gregori Labarta, and Kira Von-kleist. I am grateful for the constant support from my lab mates. We genuinely are interested in seeing each other succeed. I never take such a collaborative lab with solid labmate support for granted. The camaraderie we shared, both within and outside the lab, will always be cherished. I am thankful to my research assistants for their help in data collection, coding, and providing insightful discussions.

I would like to thank my past and present mentors and coauthors (Drs. Karen Dobkins, Upali Nanda, Renae Mantooth, Brian Hall, and Peilian Chi), who led me into research, without whom I would have never experienced the immense joy of learning knowledge and conducting research with others.

I would like to extend my gratitude to Jeff Compton for programming the testing paradigm of our studies into an iOS app. His help is invaluable as it addressed the significant

challenge of online testing we faced during COVID. I am also deeply grateful to Prof. Gail Heyman for building this connection for us.

Additionally, I am thankful to the preschool participants and coordinators at both the Early Care and Education Center (ECEC) of UCSD, as well as those in kindergartens and elementary schools in Beijing.

My fellow graduate students, particularly those in my cohort, made my transition to the US smoother. The bonding over various social events and challenges, like our statistics class, made me feel that I belong. I am thankful to my research assistants for their dedication in data collection, coding, and providing insightful discussions.

I am deeply grateful for my parent's endless support, and for the family environment that nurtured my growth mindset acuriosity for science and research. Their belief in my choices, from undergraduate studies to my Ph.D. journey, has been my backbone.

Lastly, I wish to acknowledge my partner Quynh Duong. His companionship, both personal and academic, has enriched my life in countless ways.

Chapter 1 contains unpublished material coauthored with Karen Dobkins. The dissertation author was the primary author of this chapter.

Chapter 2, in full, is a reprint of the material as it appears in Consciousness and Cognition 2023. Guo, Xinqi; Dobkins, Karen, Elsevier. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is currently under revision and is anticipated to appear in Consciousness and Cognition 2023. Guo, Xinqi; Dobkins, Karen, Elsevier. The dissertation author was the primary investigator and author of this paper.

VITA

2017 Bachelor of Social Science in Psychology, University of Macau, Macau SAR, China

2019 Master of Arts in Experimental Psychology, University of California San Diego

2023 Doctor of Philosophy in Experimental Psychology, University of California San Diego

PUBLICATIONS

Guo, X., & Dobkins, K. (2023). Private speech amount positively predicts memory performance in young adults. *Consciousness and Cognition*, 113, 103534.

The Forecasting Collaborative. Insights into the accuracy of social scientists' forecasts of societal change. *Nat Hum Behav* 7, 484–501 (2023).

Guo, X., Mantooth, R., Nanda, U., & Chilukuri, L. (in press). Establishing a Relationship between Residence Hall Design and Depression in First-Year College Students. *The Journal of College and University Student Housing.*

Hall, B. J., Xiong, P., Guo, X., Sou, E. K. L., Chou, U. I., & Shen, Z. (2018). An evaluation of a low-intensity mHealth enhanced mindfulness intervention for Chinese university students: A randomized controlled trial. *Psychiatry Research*, 270, 394-403.

CONFERENCE POSTERS AND PRESENTATIONS

Guo, X. & Dobkins, R.K. (February 2021). You are an Adult and You Still Think Out Loud? Two Empirical Studies on External Self-talk. Poster at the Annual Convention of the Society for Personality and Social Psychology

Guo, X. & Dobkins, R.K. (November 2018). Exploring Human Inner Experience by "Track Your Thoughts" – An Experience Sampling Study. Poster accepted at the 59th Psychonomic Society Annual Meeting, New Orleans, Louisiana, USA.

Guo, X., Xiong, P., Sou, E., Hall, B. J. (July 2017). An Evaluation of a Brief Mindfulness Intervention for University Students: A Pilot Randomized Controlled Trial. Talk presented at the 37th Annual Meeting of the Stress and Anxiety Research Society, Hong Kong, S.A.R., China.

ABSTRACT OF THE DISSERTATION

Speaking to Ourselves: Establishing the Cognitive Benefit of Private Speech in Young Adults

by

Xinqi Guo

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2023

Professor Karen Dobkins, Chair

Humans engage in self-talk, a phenomenon potentially unique to our species. Research indicates that inaudible self-talk ("inner speech") and audible self-talk ("private speech") together occupy up to a quarter of adults' conscious time, influencing our mental well-being as well as cognitive functioning. While inner speech has traditionally been the primary

research focus of adult cognition, its covert nature makes it hard to quantify. Therefore, most research in this area employs indirect data collection methods like self-reporting and verbal interference, which assesses the impact of self-talk by inhibiting it. However, the lack of precise quantification of self-talk prevents the accurate assessment of the strength of its effects. Private speech, another form of self-talk adults engage in, is often overlooked in research on adult cognition. However, private speech could offer unique insights into the cognitive impacts of self-talk due to its objective quantifiability.

Chapter 1 delves into whether participants' self-reports could validly inform comparisons between the impacts of inner and private speech on performance in a card-matching memory task. Specifically, I evaluated the reliability and accuracy of self-reported self-talk usage. Findings show discrepancies between objective and subjective private speech, which questions the viability of employing self-reports to precisely evaluate the effect of self-talk. Chapters 2 and 3 then focus on private speech due to its objective quantifiability and relative lack of study. In Chapter 2, participants' performance was repeatedly measured in two "Private Speech" trials, where they were instructed to complete the task (same as in Chapter 1) while using private speech extensively. We found that participants performed significantly better on trials for which they used more private speech, regardless of individuals' baseline task competency. Chapter 3 reaffirms this positive association and further establishes causality between private speech usage and enhanced performance. Performance improvements from private speech were consistent across varying task difficulties but were most pronounced in people who habitually use private speech in their everyday lives. The consistent findings across the studies underscore the potential of private speech as a tool for adult cognition, offering significant insights for educational and instructional strategies.

INTRODUCTION

Humans can talk to themselves, and we might be the only animal species capable of self-talk. In fact, we talk to ourselves quite a bit. Empirical studies show that these internal monologues or imagined conversations that are directed to no one else occupy up to a fourth of our conscious time (Hurlburt et al., 2016).

Self-talk is not just idle chatter in our heads. Empirical evidence from clinical populations underscores the significance of self-talk in our normal functioning. When an individual's self-talk goes awry, it often aligns with psychiatric, developmental, or linguistic challenges. For instance, research highlights the negative consequences of distorted self-talk in those grappling with issues like anxiety, anorexia, and depression (Morin, 2012). Furthermore, studies have delved into extreme cases where self-talk manifests as pathological voice hallucinations (Allen et al., 2007; Langdon et al., 2009). In essence, the quality of self-talk can significantly influence our mental health and well-being.

Background

Self-talk has also received extensive theoretical and empirical research interest in its potential to aid cognition. Two separate seminal theories provide the theoretical foundation that supports the cognitive importance of self-talk: Vygotsky's social-cultural theory (Vygotsky, 1987) and Baddeley's multicomponent framework of working memory (Baddeley, 2012). Vygotsky proposed a developmental trajectory of self-talk and emphasized its pivotal roles in self-regulation and internalizing knowledge. Originating from social interactions with caregivers in early childhood, self-talk evolves from external to internal communication, aiding individuals in behavioral control. For instance, a four-year-old child might say to herself, “Lily, don’t touch the toaster oven.” when alone, repeating and internalizing a rule that her mother told her earlier. Vygotsky's insights have profoundly influenced research areas like cognitive development (see

Winsler, 2009 for a review), sports psychology (see Hatzigeorgiadis et al., 2011 for meta-analysis), and second language acquisition (see Guerrero, 2018 for a review). Baddeley's multicomponent framework has been applied to various cognitive activities and holds significant influence in cognitive psychology literature (see Baddeley, 2012 for a review). This framework suggests that working memory comprises a central executive function that is in charge of tasks like focusing attention and task switching. This central function governs three subordinate systems. One of these, the phonological loop, is particularly relevant to self-talk. The phonological loop stores phonological or auditory information, and it has two parts: the phonological store that retains the auditory information we hear ("inner ear"), and the articulatory loop that facilitates rehearsal or repetition ("inner voice"). A classic example of the articulatory loop is when we repeat a phone number until we can jot it down. Baddeley's model sheds light on the potential interactions between self-talk and other working memory components, especially through verbal recoding and rehearsal, emphasizing its role in memory retention. Further, this framework has popularized the method of using verbal interference to probe the influence of self-talk on cognition. I will delve deeper into this in the section titled "Challenges in Quantifying Self-Talk and Its Cognitive Impacts".

I draw high-level inspiration from both Vygotsky's and Baddeley's frameworks, as well as the studies they have influenced, to suggest a promising approach for researching self-talk's impact on adult cognition. My specific takeaways from these frameworks are as follows: Vygotsky's insights into self-talk development emphasize the relevance of audible self-talk (i.e., private speech) in cognition, in addition to the more commonly researched inaudible self-talk (i.e., inner speech). As will be discussed below, private speech has a salient methodological advantage to being a vehicle for studying self-talk. Complementing the theoretical approach of

Vgotsky, research rooted in Baddeley's framework provides specific insights into the mechanisms of how self-talk interacts with other sub-components to assist working memory.

Challenges in Quantifying Self-Talk and Its Cognitive Impacts

Despite the extensive research interest in self-talk, the majority of existing studies primarily investigate which cognitive functions language may or may not influence but often do not address the magnitude of these effects. This limitation is partly because of the methodological challenges in data collection, particularly when it comes to quantifying self-talk usage (Alderson-Day et al., 2015; Guerrero, 2018). Quantifying adults' self-talk usage presented unique challenges due to its (frequently presumed) covert nature. As a result, researchers either avoid quantifying self-talk usage or use indirect methods of data collection.

A solution to the methodological challenge of quantifying self-talk is through self-reports. However, one prevalent critique of this approach is recall bias. Instead of reflecting on the entire task duration, people often base their reports on the most recent and notable experiences during the task, such as successes or failures. Indeed, the often condensed and fragmented nature of inner speech complicates its quantification, especially during concurrent tasks. This difficulty is further exacerbated by the absence of signals such as auditory input and articulatory output. As a result, the usefulness of self-reported inner speech usage[1] has been questioned due to its lack of convergent validity – measures of similar self-talk usage do not correlate well (Uttl et al., 2011); further, respondents struggle to recall characteristics of their self-talk that might be important for quantification, like whether it was brief and fragmented or expanded and similar to conversational language (McCarthy-Jones & Fernyhough, 2011).

[1] Despite the criticism on its ability to quantify self-talk, self-report tools have shown strong statistical reliability and are insightful for understanding the language choice, content, and functions of self-talk (Uttl et al., 2011).

Apart from self-reports, articulatory suppression, a frequently used verbal interference technique, examines the effects of self-talk by inhibiting it. Essentially, this method involves repeatedly articulating a short, irrelevant word or syllable out loud during a task. Studies employing this method suggest that (internal) verbalization could influence cognition by imposing labeled categories from long-term memory to carve up the continuous stimuli space, such as colors and shapes. For example, Souza and Skóra (2017) reported a detrimental effect of articulatory suppression on the recall accuracy of colors, as measured by reproducing them on a continuous color wheel. Overtly labeling, in contrast, was found to facilitate recall accuracy but also made the memory representation more categorical. Nakabayashi and Burton (2008) found that articulatory suppression during encoding harms facial recognition memory when compared with both a verbalization condition (where participants were asked to describe the faces out loud during encoding) and a tapping control condition. Interestingly, the timing of verbalization matters. The same study's Experiment 4 showed a weak detrimental effect of verbally describing the faces five minutes *after* the visual presentation. This might be attributed to verbal descriptions becoming more generic at that point, unlike during encoding or immediately after visual presentation, when they include more diagnostic facial features (Wilson et al., 2018). Inhibiting self-talk has been shown to decrease inhibitory control. For instance, participants under articulatory suppression made more perseveration errors during the Wisconsin Card Sorting Task (Dunbar & Sussman, 1995). Similarly, verbal interference was found to result in more impulsive responses in a Go/No-Go task (Tullett & Inzlicht, 2010).

Although articulatory suppression is useful for determining the role of self-talk in specific cognitive processes, the quantity of self-talk during non-suppressed conditions remains elusive. Such knowledge is crucial to understanding the relative strength of self-talk across various

cognitive tasks. For instance, minimal self-talk could greatly improve performance in task A if its effect is potent. Conversely, self-talk might have lower efficacy in task B. But if it is more frequent in task B, the overall benefits in tasks A and B could seem comparable, masking the true differences in their effectiveness.

Therefore, obtaining high-quality data on self-talk usage is essential for understanding its influence on cognition. Unfortunately, the imprecision of self-reported self-talk metrics, paired with a lack of effective methods to objectively measure covert self-talk, limits accurate evaluations of its impact on cognitive tasks.

Private Speech in Adults: Overlooked Potential and Existing Research Gaps

Human adults actually engage in two forms of self-talk: covert/inaudible self-talk ("inner speech") and overt/audible self-talk ("private speech"). I argue that private speech offers a promising avenue to explore the cognitive impact of self-talk in adulthood because of its objective quantifiability and its understudied status. Studying private speech can complement ongoing research on self-talk's role in adult cognition, which primarily focuses on inner speech.

There exists a limited array of studies examining the effect of private speech on adult cognition. The majority are either correlational or analyze shifts in the frequency and nature of private speech under varying task demands. Duncan and Cheyne (2001), for instance, highlighted that adults spontaneously engage in less private speech during motor tasks (e.g., paper folding) compared to verbal tasks (e.g., digit entering, similar to transferring phone numbers). However, their study did not elucidate whether private speech enhanced performance or merely co-occurred. Adults have also been observed to engage in private speech during challenging cognitive tasks (Alarcón-Rubio et al., 2013; R. M. Duncan & Cheyne, 2001), while mastering new manual skills, such as crafting lanyards (Soskin & John, 1963), and in

embarrassing social situations (Duncan & Tarulli, 2009). Current studies on adult private speech seem to view it more as an accompanying behavior during concurrent activities rather than a behavior that can influence cognition. Such an approach, while informative, is insufficient, as it leaves the potential active role of private speech in adult cognition largely unexplored.

The limited research on adult private speech, as mentioned earlier, stems in part from Vygotsky's dominant theory on self-talk development, which suggests a temporary role of private speech in cognition. According to Vygotsky, private speech is a developmental precursor to inner speech, which is a more mature form of self-talk that is used more frequently after middle childhood. Thus, the Vygotskian theory regards inner speech as the end result of a gradual internalization of the child's self-directed language, wherein private speech exists temporarily before children's inner speech reaches maturity. Empirical studies do, in fact, support the theory in showing that the frequency of private speech peaks during the preschool period (around 3-6 years of age), after which private speech gradually becomes more inaudible (Berk, 1986; Winsler et al., 2003). Perhaps because of the prevalent belief that the frequency and significance of private speech decline before adulthood, its role in adult cognition remains largely unexamined, especially when compared to the extensive studies on children's private speech and adults' inner speech.

In sports psychology and second-language learning studies, self-talk research targets a broad age range: from young children to adults of various ages. These fields typically examine self-talk's impact on performance by encouraging participants to use it. However, many studies in these fields either do not monitor the actual amount of self-talk used or depend on participant self-reports, the limitations of which were previously discussed. Additionally, these studies often do not differentiate between inner speech and private speech.

Production Effect: Memory Enhancement Through Vocalization and Its Methodological Challenges

The production effect in cognitive psychology explores a phenomenon in adults closely related to private speech. This line of research investigates the differences between externalized encoding methods (reading aloud, typing, singing) and internalized encoding methods (e.g., silent reading) on the memory of the studied materials. Essentially, the production effect suggests that information read aloud is better remembered than when studied silently (MacLeod et al., 2010). A typical experiment of the production effect might present participants with words on a computer screen, color-coded for distinct encoding methods, for instance, blue for reading aloud and black for reading silently. Subsequent recognition tests often indicate a superior performance for the words read out loud, over the words read in silence. This memory advantage is often explained through the *distinctiveness* account (MacLeod & Bodner, 2017), which proposes that reading out loud encodes material better in memory due to the additional involvement of articulation (like the physical motions of the tongue and throat) and the auditory feedback of hearing oneself. In line with this view, the memory advantage was nullified when the verbalization was non-distinct, as when participants merely responded uniformly with the word "yes" to all items during encoding. Further, Richler, Palmeri, and Gauthier (2013) found that naming images of examples from just two categories ("chair" and "lamp") did not improve recognition. In some cases, it even had a detrimental effect, likely due to the lack of distinctiveness during the encoding process. This indicates that simply vocalizing or producing a sound is insufficient. Instead, the distinctiveness, in other words, the variety and uniqueness, of the auditory and articulatory experience matter in the encoding process.

However, the production effect literature has a methodological issue that mirrors the concern raised in research approaches like articulatory suppression. While these studies do employ a control group for silent encoding, the extent of inner speech during this control condition is not rigorously quantified, which raises questions that point to various alternative explanations: Were participants truly engaging in silent reading in the control condition, or did they also rely on visual means to encode the words? Might irrelevant inner speech occur during the silent encoding control, unbeknownst to the researchers, potentially diminish the efficacy of silent encoding and exaggerate the effect of reading aloud? Thus, without robust approaches to quantify inner speech, it's unclear if the memory advantage of the aloud over the silent reading condition stems from a lower extent of silent encoding or if silent reading is indeed less effective than its aloud counterpart.

Despite its methodological limitation and the relatively restricted content (i.e., pre-determined materials, like words) for verbalization, findings from the production effect research suggest that prompting private speech might facilitate cognition as effectively as, or perhaps more than, encouraging inner speech.

Potential mechanisms on how private speech might improve performance in the context of a card-matching memory game

While we are interested in the role of private speech in adult cognition, our inferences are largely shaped by the testing paradigm in which participants engage. In this dissertation, I examine the impact of self-talk on cognitive performance with a card-matching memory task as the primary testing paradigm. While other researchers might employ tasks like the Tower of London or Dimensional card sorting – which are executive functioning tasks and not predominantly memory-based – to explore the effects of private speech, the card-matching task

offers notable advantages. Its highly straightforward rules, engaging nature, and accessibility make it suitable for a wide range of demographics, such as young children (Eskritt & Lee, 2002; Schumann-Hengsteler, 1996), and even non-human primates (Martin & Shumaker, 2022; Washburn & Gulledge, 2002).

Drawing from extensive adult self-talk studies, which often focus on inner speech, and the distinctiveness account that explains the production effect (contrasting external and internal encoding strategies), I propose several non-mutually exclusive mechanisms suggesting that private speech is beneficial for cognitive performance within the context of a memory task.

First, private speech can activate long-term conceptual knowledge to enhance working memory. Preliminary internal testing and prior research have shown that adults spontaneously name and describe visual stimuli unless inhibited or instructed otherwise. Labeling and recoding visual stimuli into verbal descriptions can impact memory representation, a finding supported by visual working memory research. Second, private speech augments the distinctiveness of the studied material. Beyond the act of verbalizing visual stimuli, private speech enhances the distinctiveness of studied material via auditory and articulatory signals, an idea rooted in the production effect literature. Further, private speech bolsters motor control through heightened attention. The idea that self-talk aids motor control is grounded in Vygotskian self-regulation theory. Here, our inner voice, akin to hearing guidance from another individual, helps regulate our actions and concentrates attention on motor control and can cue specific subcomponent motor actions, enhancing the overall movement goal. Lastly, from a neuroscience perspective, private speech can improve performance through the general higher involvement of neuro activities. This idea can be tested through brain imaging studies, to see if areas representing the images show heightened activity when talking out loud vs. being silent.

Current Direction

The functional role of self-talk in cognitive functioning has been a subject of research for a long time. However, several limitations in the existing literature prevent a deeper understanding of self-talk's impact on cognition. Firstly, many studies do not monitor or quantify self-talk, making it hard to determine the true magnitude of its effects. Secondly, when self-talk is quantified via self-reports, the accuracy of these reports remains uncertain. Thirdly, while a few studies focus on audible private speech during adulthood, most investigate spontaneous private speech, leading primarily to correlational conclusions about its relationship with performance.

This dissertation examines how private speech, a form of self-talk, affects the cognitive performance of young adults. It comprises three chapters, wherein participants were prompted to use private speech during a card-matching memory task. By leveraging the objective quantifiability of private speech, this dissertation provides a more precise and reliable evaluation of the effect of self-talk. Overall, the findings emphasize the relevance and potential benefits of private speech in adult cognitive processes, even after the presumed maturation of inner speech.

In Chapter 1, I discussed a persistent methodological challenge in studying the role of self-talk in cognition: while private speech usage can be objectively quantified, inner speech largely relies on self-reports, the validity of which has been shown to be questionable. To address this, we created a framework to assess the truthfulness of self-reports, particularly when participants were in a concurrent cognitive task. Our aim was to determine the accuracy and consistency of self-reported self-talk usage. The findings show that while self-reported self-talk usage showed high reliability, they did not align with objective metrics (i.e., subjective and objective private speech metrics were not correlated). This adds to the existing literature that questions the trustworthiness of self-reported self-talk usage. Therefore, we steered subsequent

research direction towards private speech, an under-researched yet methodologically tractable form of self-talk.

In Chapter 2, I started to specifically investigate the functional role of private speech on cognition by probing a within-person association between concurrent private speech usage and cognitive performance. Findings indicate better performance in trials when participants use private speech more frequently. Further, this effect was also consistent irrespective of participants' baseline competencies. This chapter establishes that, among young adults, increased private speech usage correlates with enhanced cognitive performance.

In Chapter 3, I explored the positive association in greater depth, aiming not only to replicate it but also to test for causality. The results indicated that when participants were prompted to use private speech, their task performance improved compared to when they were instructed to remain silent. The advantages of private speech were consistent across different task difficulties. Notably, individuals who regularly engaged in self-management through private speech in their everyday lives experienced the most benefits from the private speech manipulation. This chapter strengthens our belief in the positive impact of private speech on cognitive performance among young adults.

In summary, this dissertation contributes to the literature by using a more tractable form of self-talk, offering good-quality data to elucidate the associations between private speech, cognitive performance, and individual differences in young adults. The subsequent chapters delve into the discussions and implications of the findings in detail, offering valuable insights for both theoretical understanding and practical applications in educational contexts. I suggest that private speech warrants further attention in research and theory development. The current theory

on self-talk development may need refinement, especially to clarify how private speech can benefit cognition in adulthood.

Overall, I believe that insights from private speech present a promising approach to deepening our understanding of self-talk's impact on cognitive performance in adults.

CHAPTER 1 ASSESSING THE RELIABILITY AND ACCURACY OF SELF-REPORTED MEASURES IN SELF-TALK USAGE

Abstract

Adults engage in two types of self-talk: inner speech (inaudible) and private speech (audible). While inner speech's effect on adult cognition is well-studied, private speech's role and its efficacy on adult cognitive performance relative to inner speech remain largely unexplored. A precise evaluation of these effects necessitates good-quality metrics for private speech and inner speech usage. While private speech usage allows for both objective measurements and self-reporting, quantifying concurrent inner speech usage relies on self-reports due to its inaudible nature. In response, we formulated a framework to evaluate the quality of self-reported self-talk usage: we expect strong correlations between self-report and objective measures ("accuracy") and consistency across different self-report scales ("reliability"). If self-reports meet the accuracy and reliability criteria, we can then confidently employ self-reports and start to compare the effectiveness between inner speech and private speech. Otherwise, private speech would be my focus in assessing the influence of self-talk on adults' cognition because of its objective quantifiability and the understudied nature.

In a study based on this framework, adult participants engaged in either private or inner speech during a card-matching task. They then reported their self-talk usage using percentage and Likert scales. For private speech only, we also attained an objective metric through audio recording. Findings showed proof of reliability, but a lack of correlation between the subjective and objective private speech metrics. Hence, we move forward with an emphasis on private speech. Strategies to enhance the accuracy of subjective self-talk usage reports were discussed.

Introduction

The COVID-19 pandemic disrupted my dissertation studies, making it challenging to conduct proper laboratory experiments. However, this period provided an opportunity to pilot some pressing questions that would guide our future decision-making. One decision to make is whether self-reports truthfully reflect actual self-talk usage during a cognitive task and therefore can be used to precisely evaluate the effect of inner speech on cognition. If self-reports proved unreliable, however, my focus in this dissertation would shift exclusively to private speech due to its methodological advantage of being a tractable behavior.

Pilot Study Overview

In the pilot study, participants were instructed to engage in private speech (in the PS condition) and inner speech (in the IS condition) during a memory task that would be used in subsequent studies of this dissertation. Participants reported the amounts of inner speech in the IS condition and private speech in the PS condition. For an objective measure, we audio-recorded participants during the PS trials. The primary aim was to gauge the reliability and accuracy of these self-reported amounts.

Criteria for the Trustworthiness of Self-reported Self-talk

If self-reported usage is a reliable and accurate measure of self-talk, there should be:

1. A strong correlation between objective metrics of private speech (from audio recordings) and subjective reports.
2. A high "test-retest reliability" of the same behavior in different question formats. Due to a lack of established measurement of self-talk usage during a concurrent task, we designed in-house self-reports in two formats (percentage scale and Likert scale) for both inner speech and private speech. The expectation here is that inner speech ratings on the percentage scale will correlate strongly with their corresponding Likert scale ratings.

Similarly, private speech ratings on the percentage scale should be highly correlated with their Likert scale counterparts.

To preview our findings, although the self-reported usage correlated strongly among themselves, self-reported private speech did not correlate with audio-recorded private speech metrics. The discrepancy between self-reported and objective private speech amounts makes us skeptical about self-reported inner speech's accuracy. Given that private speech's subjective assessment benefits from external cues like auditory input, and inner speech doesn't, it's challenging to rely solely on self-reports. This skepticism impacts our confidence in using such metrics for our main research questions.

Thus, while preliminary data suggests self-reported self-talk reliability, it doesn't necessarily depict actual engagement. Ways to effectively utilize these self-report measures will be elaborated upon in the Discussion.

Method

Participants

The participants of this study were undergraduate students who were recruited from the participant pool managed by the Department of Psychology at the University of California San Diego. Recruitment took place between January and November of 2021, under the remote-testing restriction due to COVID-19. To be eligible for the study, the individuals must be 18 years or older and own an iOS device. The latter requirement was due to our testing paradigm being reliant on a specifically developed iOS app, which was created by a professional as a solution to circumvent the restrictions of in-person testing due to the pandemic. Participants were compensated with course credits for their participation. This study was approved by the Institutional Review Board at our university.

The sample consisted of 43 participants. As detailed in the study design (below), each participant engaged in four behavioral trials. The repeated testing approach strengthened our confidence in having adequate power to answer our simple research questions (i.e., reliability and accuracy of self-reported usage of self-talk). Participants' ages ranged from 18 to 35 years (M = 20.46, SD = 2.64), and their gender identities were 63.9% female and 36.1% male.

Apparatus and Material

Card-Matching Task

The study used a card-matching game called "Concentration", 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 (two cards were tapped and revealed at a time).

In the current study, we used the card-matching game in a 5 × 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 × 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).

Creating Stimuli for the Card-Matching Task

Prior to the pandemic, we intended to extend our testing paradigm to children and wanted to employ images that were labelable by both children and adults. To this end, we first 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). In our selection process, we opted for

concrete nouns from Word Bank that can be produced by at least 65% of 30-month-olds[2]. This criterion was based on the assumption that such nouns could be produced by all 4- to 6-year-olds (which was one of our original target age groups). Once we identified the suitable nouns, we then searched corresponding clip-art images via Google. The clip-art patterns were all in color with white backgrounds.

Self-Reported Usage of Self-Talk

The same set of self-talk usage questions was asked immediately after each of the actual trials: with the majority of the text shared between the inner speech condition and private speech conditions (details in Procedure), and the only difference being the modes of self-talk the participants were instructed to engage in during the upcoming trial.

> *Inner Speech - Percentage:* We realize we asked you to *talk to yourself internally as much as you can* during the game, but still, people differ in how much they do this. With this in mind, please let us know…..during the last trial, on a scale of 0-100, what *percentage* of your thought was words/language versus any other types of thoughts (e.g. visual imagery, music, abstract, or nothing)?
>
> *Private Speech - Percentage*: 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. With this in mind, please let us know ….during the game, what *percentage* of the time were you talking out loud to yourself (as opposed to being silent)?
>
> *Inner Speech and Private Speech Usage - Likert Rating* (see Introduction for the rationale of assessing the self-reported quantity of self-talk in two ways): The same text was used with the respective mode of self-talk items above, but Likert rating questions asked about

[2]At the time the data was downloaded (February 19, 2020), Word Bank was monitoring the word production of children up to the age of 30 months.

the extent of self-talk on a 7-point Likert scale. This Likert scale has the lowest point and the highest point labeled “Not at all”, and “Completely/Entirely”, respectively.

Procedure

More information, like the minor procedural details specific to online testing over Zoom, can be found in Appendix A, immediately after this chapter.

Asynchronous online preparation

Prior to the online testing session, participants received a preparation email containing detailed instructions for the preparation of the experiment. The email included the following guidelines:

Downloading the App: Participants were instructed to download the “Concentration Cat” app from the App Store on their iOS devices.

Video and Audio Settings: Participants were asked to turn off their cameras and unmute themselves to allow audio recording during the actual trials. Participants were asked to turn on their cameras while the experimenter delivered instructions, to ensure accountability of participants’ attentiveness.

Be in a Quiet Space: Participants were asked to find a quiet and secluded environment, where their performance would not be interrupted during the testing session to minimize external distractions that could impact their performance.

Consent Forms: Participants were requested to fill out consent forms for general study purposes and for audio recording before the start of the testing session. This means that participants knew they would be recorded during the task beforehand.

Synchronous Online Testing

The experimenter informed the participants that they would be playing a card-matching game, which was explained to them through a pre-recorded video demonstration played through

the screen-share function over Zoom on the experimenter's computer. The video demonstration featured a 2 × 3 array of face-down cards with patterns different from those used in the actual trials. During the video, the experimenter paused periodically to elaborate on the rules and goals of the game. Next, the experimenter proceeded by instructing participants to play four actual trials of the game on their own mobile devices, through the Concentration Cat app. During the actual trials, the experimenter turned off their own camera and clicked "leave audio" of Zoom, which temporarily blocked the experimenter's access to the participants' audio, so as to not make the participant uncomfortable in engaging in self-talk. The experimenter only opened the camera and joined the audio in between the trials to deliver instructions for the next trial. The four actual trials were presented in a pre-designed order, as follows.

Self-talk conditions were counterbalanced across participants, with half of the participants starting with the Private Speech (PS) condition, and the other half starting with the Inner Speech (IS) condition. To have a more reliable evaluation, we had two trials within each speech condition. Further, to control for order effects[3], we adopted an "ABBA" design: Each participant would go through one of the two speech conditions (A), then the other speech condition (B), then the other speech condition again (B), and finally the first speech condition (A) for the second time. This resulted in participants being randomly assigned into one of the two different orders of the four trials:

PS – IS – IS – PS

IS – PS – PS – IS

[3] Changes in performance that occur not because of the condition differences, but because of their order. For instance, participants may perform better in the second condition simply because they have had a chance to practice, or worse because of fatigue.

In the IS trials, participants were asked to finish the game in as few turns as possible, and were instructed to engage in inner speech throughout the task. Specifically, they were told, "*Please finish the game using as few taps as you can. You will see the time and taps you used after each trial. But the only goal is to use as few taps as you can to finish the game. We do not care about the time taken to finish the game in this study. You can finish the game at your own pace. While you're playing the game, talk to yourself internally or in your head as much as you can throughout the game. You can use the language you're comfortable with. We do not have instructions on the content of your self-talk. I (the experimenter) will leave the Zoom audio during your actual trials, and I will ask you to turn off the camera, so I won't be able to see you or hear you during the game.*". In the Private Speech trials, participants were given similar instructions but were asked to talk out loud instead. Specifically, they were told: "*Please finish the game using as few taps as you can. You will see the time and taps you used after each trial. But the only goal is to use as few taps as you can to finish the game, and we do not care about the time taken to finish the game in this study. You can finish the game at your own pace. Talk to yourself audibly or externally throughout the game or as much as you can. You can use the 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 of your social conversations. I (the experimenter) will leave the Zoom audio during your actual trials, and I will ask you to turn off the camera, so I won't be able to see you or hear you during the game.*".

We recorded participants' audio through Zoom and later used it to calculate an objective measurement of the amount of their private speech (see *below*). The audio recordings were collected for all four trials (the IS and PS trials). After each actual trial, the participants were instructed to report their self-talk usage of the trial that they just finished.

This process was repeated for each of the four actual trials. The participants' demographics were asked at the end of the study.

Amount of Private Speech (PS).

The current study used utterances/minute as the metric for the actual usage of private speech. The choice of this metric (as opposed to *total* utterances or metrics based on the number of *words*) is justified in Guo & Dobkins (2023) or Chapter 2, noting that it has also been used in previous private speech studies (Duncan & Cheyne, 2001; Fernyhough & Fradley, 2005; Kronk, 1994; Mulvihill et al., 2021). As a first step, the audio recordings of participants' private speech were analyzed offline by the first author and her research assistants. Next, data were entered into a spreadsheet in units of "utterances", defined as an audible verbal unit separated by differences in semantic meaning or at least one second of temporal distance (Frausel et al., 2020; Rowe, 2012; Rowe & Goldin-Meadow, 2009). For example, "Dog at the top right corner" would be considered as one utterance, whereas "Is the dog here? Nope." would be considered as two utterances. As a final step, utterances/minute was calculated as the number of utterances divided by the time to finish the trial.

Results

Descriptive Result

Descriptive data of means and distributions of study variables are presented from 172 trials (4 trials × 43 participants). For the Private Speech trials, the mean number of utterances/minute was 37.01 (*SD* = 14.52), which was higher than the values observed in Chapter 2 (*M* = 27.56, *SD* = 11.26) and Chapter 3 (*M* = 27.56, *SD* = 11.26), which also employed the same memory task with easy-to-label images[4]. The difference in the amount of private speech

[4] Chapter 3 also has a condition with hard-to-label images that lead to lower amount of private speech utterances. Details see Chapter 3.

recorded might be due to methodological differences between the current chapter and the rest of the dissertation, for instance, only the participants of the current study were tested online and were informed in advance that they would be audially recorded.

The mean ratings of the self-reported percentage and Likert rating of *private speech* were 77.21% (*SD* = 22.43%) and 5.65 out of 7 (*SD* = 1.32), respectively. With regard to the self-reported amount of *inner speech*, its percentage (*M* = 70.50%, *SD* = 26.07%) and the Likert rating (*M* = 5.08, *SD* = 1.63) seems to be lower than their private speech counterparts[5].

Are People Good at Reporting Their Actual Usage of Private Speech?

As a first step, we asked whether the two types of self-reported amount of private speech (Extent and Percentage) were associated with each other. Using a Type III sum of squares multilevel regression model for the PS condition, the dependent variable was Likert Rating and the predictor term was Percentage, with Participant included as a random intercept effect. Having the percentage scale rating as the predictor is more intuitive as the percentage scale items were asked prior to the Likert items. Because the two were found to be significantly and strongly associated (β = 0.80, 95% CI = [0.67, 0.93], $p < 0.001$), this suggests that the subjective measure for private speech is quite reliable.

Next, using a Type III multilevel modeling, we asked how well do self-reported metrics (the predictor variable) reflect the objective private speech usage (the dependent variable), with participant entered as a random intercept effect. The results show a lack of association between self-reported private speech and the actual amount of private speech, no matter which self-report

[5] Paired-sample t-tests revealed a significantly higher subjective amount of private speech in the PS condition than inner speech in the IS condition, the t-tests for both percentage and extent had *p*s < 0.05. We did not hypothesize this difference, and it is out of the scope of the dissertation to explain why the self-talk usage was different when both PS and IS conditions had the instruction of engaging in the respective form self-talk as much as possible. Future studies that attempt to contrast inner speech and private speech should take these results into consideration.

format was used (Percentage: β = 0.14, 95% CI = [-0.06 – 0.41], p = 0.213; Likert Scale: β = 0.17, 95% CI = [-0.06 – 0.41], p = 0.140). This suggests that self-report measure was not a trustworthy measure of the actual private speech usage. One possibility is that this weak association results from low reliability in one or both of the measures. However, we believe this explanation is unlikely, since the high correlation between Likert and Percentage ratings suggests good reliability for the subjective measure, and inter-rater tests suggest good reliability[6] in the coding of the actual usage data..

Since the result did not meet the first of the two criteria we set in the intro to deem self-reported measurement trustworthy of reflecting the actual self-talk usage. We could have stopped here and moved on to actual usage metrics. However, we still deem it informative to understand the reliability of the inner speech metric, which is presented below.

Is the Subjective Inner Speech Usage Report Reliable?

Mirroring the analyses performed on self-reported private speech above, we asked whether the Likert ratings and Percentage of inner speech were associated with each other. We used a Type III sum of squares multilevel regression model, the dependent variable was Inner Speech Likert and the predictor term Inner Speech Percentage, with Participant included as a random intercept effect. The results revealed that the two were significantly and strongly associated (β = 0.89, 95% CI = [0.79, 0.99], $p < 0.001$), which indicates that the self-reported inner speech measures are statistically reliable.

Discussion

Although we identified a strong correlation between self-reported amounts of self-talk on percentage and Likert scales, the meaningfulness of this correlation may be questioned: the two

[6] Due to researcher availability, there was not an additional coder for data in this chapter, but see Chapters *2 and 3* for proof of generally high inter-rater reliability of private speech quantity and content.

self-reported questions were placed next to each other on the questionnaire after each trial, which could have led participants to perceive them as duplicate queries. Consequently, they might have simply converted their response from one format (percentage scale) to another (Likert scale), instead of treating the two questions as unique items that require equal attention. Furthermore, our self-reporting method failed to meet a crucial criterion: for self-talk types that can be objectively quantified (i.e., private speech), there should be a significant positive correlation between the self-reported and objective amount of self-talk.

Since previous studies have shown that children as young as 8 years old display adult-like introspection of their own internal processes (Flavell et al., 1993), this makes it somewhat surprising to find no association between young adults' objective usage of private speech and its subjective counterpart. I think this result can be reconciled with previous findings. Although young adults can reflect on the *presence* of their self-talk (e.g., asked in a binary "yes" or "no" format), they may struggle to accurately gauge its quantity. Further, the discrepancy between objectively measured and self-reported private speech may also indicate that they are measuring different constructs. Participants might interpret self-reported measures as relative to their typical behavior, thus biasing their responses. For instance, knowing that they were instructed to talk aloud as much as possible might lead participants to inflate their rating on self-talk usage, leading to an overall high rating, despite our efforts to frame the question objectively (we ask them about the "proportion of time" they engaged in self-talk).

In sum, the results make us question the accuracy of self-reported self-talk quantity during the cognitive task, which is especially concerning if we want to understand the within-person (Level 1) effects of self-talk usage on performance in a multilevel modeling approach

(Chapters 2 and 3). Given these potential pitfalls of self-reported metrics on self-talk usage, we decided to move forward with a focus on private speech due to its objective quantifiability.

There is a pressing need for further research and the development of methods to quantify inner speech. One method to mitigate recall bias in self-talk reporting is the Experience Sampling Method (ESM). This technique prompts participants to report on their internal processes at random intervals while they are performing a task or go about their days. Unlike other methods that rely on a single summary item for the entire task, ESM aggregates multiple experience points. For example, participants might be randomly interrupted several times during a trial and asked if they had just engaged in inner speech. The final rating for that trial would be based on the proportion of "yes" responses relative to the total number of interruptions. By assessing internal processes in real-time, ESM minimizes memory biases, such as recency or salience biases, especially when evaluating fleeting processes like inner speech. One downside of this approach, however, is that the interruptions might disrupt task completions and the inner speech behavior itself.

Acknowledgments

Chapter 1 contains unpublished material coauthored with Karen Dobkins. The dissertation author was the primary author of this chapter.

Appendix A

Self-talk Experiment Script

Before the testing, there'd be a preparation email sent to the participant, which contains:

A. Downloading the Concentration Cat app from the App Store

    1. Make sure to "Allow access to all photos".

B. Turn off the video during the actual tests, and unmute yourself

C. Find a quiet space where your performance will not be interrupted.

D. Sending the two consent forms and asking them to fill them out before the testing session starts.

E. Ask for the participants' iOS phone number/iCloud email. We will need this for sharing the testing material with the participants.

1. **Greetings & double-check their name**
    a. "Hello! Thanks for joining the Zoom session. I'm (your name) and I will be the experimenter today. Are you (the participant's name from SONA)? "
2. **Consent form and audio**
    a. "You were asked to complete two consent forms in a preparation email from us. One is for participation, and the other is for audio recording. Have you signed both of them on Qualtrics questionnaires?"
    b. "We will record your audio during the **actual testing**. Please keep your camera on and unmute yourself before that. We'll ask you to turn off the camera during the

actual testing. Please try to find a separate and quiet room with stable WiFi so that your performance won't be interrupted".

3. **Preparation**

    a. App - "Just to confirm: Have you downloaded the Concentration Cat app? Did you allow the app to access all of your photos?This is just for experiment purposes and your privacy won't be compromised, you can always restrict access after the experiment! "

    b. Shared albums - "Have you turned on the Shared Album button on your Photos App?"

    c. Pictures - "In the preparation email you were informed that we'll need your phone number to share with you the picture material needed for the study. Could you tell me your phone number, and I'll add you as a viewer of the shared albums?"

    d. The experimenter add the participants' phone number to all the four shared albums, and say, "Please accept the invitation to the shared albums named, Trial #1, #2, #3, and #4"

    e. Qualtrics – (paste this link into the chat (LINKS ARE UPDATED GO TO THE TOP OF THE DOCUMENT): "Please open the link I posted in the chat as part of the preparation. Do not proceed to the next page until I ask you to". "Please leave the Qualtrics link open throughout the testing (to prevent reopen/reload). Now please come back to Zoom."

4. **Demo trail**

a. “There will be 1 demonstration and 4 actual trials for this study. The demo is a short video clip to show you how to play the game, and it’s a simpler version of the actual trial.”

b. “The idea of this memory game is to efficiently collect all the hidden pairs. You will start the game by tapping a card from this array of cards to reveal its pattern. And when you tap another card, the patterns between the cards will be compared: if they are identical, then both cards will be removed automatically. However, if their patterns mismatch with each other, both cards will be flipped back down. You will need to keep tapping other cards in the array to finish collecting all the hidden pairs.”

c. Start sharing your screen and play the video after you finish explaining the game.

d. “As you’d expect, you’ll need to remember where certain cards are to finish the game efficiently. That’s the end of the Demo trial.”

e. “Be prepared to tell the experimenter the time and number of taps you use to finish each trial.”

5. Actual trials

a. “Now please go to the Shared Album named Trial 1 from your Photos App. There should be 12 clip-art images. Download all 12 images to your local Photo Album.”

b. “Go to the Concentration Cat app. *Long Press (or press for 2 seconds)* the All photos Album, then tap ‘Select More Photos’ from the window that pops out. Choose the 12 images you imported from Trial 1. **Do not click ‘play now with 12**

photos' yet. Only tap the button after I say you can start. Please show me the screen of your phone after you finish selecting the images".

c. (after they show you the right images being selected) "The instruction for the first trial is that:

   - **Encourage Private Speech (EPS)**: "while you're playing the game, talk to yourself externally or audibly as much as you can throughout the game. You can say anything you want. Just try to finish the game as efficiently as you can while you talk to yourself out-loud"
   - **Encourage Inner Speech (EIS):** "while you're playing the game, talk to yourself internally or in your head as much as you can throughout the game. You can say anything you want. Just try to finish the game as efficiently as you can while you talk to yourself in your head"

d. **"Don't tap the play button yet, there is one more instruction**: I will start recording the audio soon. When I do, I'll let you know. At the same time, I want you to turn off the camera and keep yourself unmuted. After I start the recording, I will leave the audio, so that I cannot hear what you're saying. After you finish this trial, please type in the chat to let me know that you're done. Is this instruction clear?"

e. "Great. I'll start the recording and leave the audio now. You can click the play button and start the game whenever you're ready. Remember to talk out-loud/talk internally"

f. (Experimenter turns off video, start the recording, double-check the participants' video is off and audio is on, and click Leave Audio)

g. (Experimenter rejoin audio whenever you see the participant's signal in the chat)

h. **"Could you show me your phone's screen so that I know the time and number of taps you used."**

i. Record the performance on the excel spreadsheet. "You can keep the game app as it is for now"

j. Direct them to the Qualtrics "Now please go to the Qualtrics link you opened in your browser and answer the first four questions of the questionnaire. The answers to this survey are just for you only, and your privacy is being protected."

---------- Below are abbreviated script due to repetition

k. "This is the end of trial #1."

l. "Please **delete** the images of Trial #1 from your local album. Go to the shared albums #2 and download the images to your local album".

m. "Now go to the Concentration Cat app, tap the Change Photos button at the bottom right of your screen."

n. "Please long press All Photos and let the Concentration Cat select all 12 images from Trial #2. Do not click the 'play the game with 12 photos yet'. I'll let you know when you can start".

o. …

CHAPTER 2 PRIVATE SPEECH AMOUNT POSITIVELY PREDICTS MEMORY PERFORMANCE IN YOUNG ADULTS

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.

Introduction

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 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 the 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[7]. 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 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

[7] 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.

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* (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)[8]. 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 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

[8] 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.

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 participants, 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 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[9].

[9] 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.

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 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 Appendix B.

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[10].

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

[10] 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[11]. 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

[11] 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.

(including categories such as “rehearsing” or “labeling”, 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[12]. 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.

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

[12] 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: $p = 0.003$).

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[13]. Across the entire 233 useable Private Speech trials, the mean number of utterances/minute was 27.56 (*SD* = 11.26). In Table 2.1, 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).

[13] 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 2.1: 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

Variables	First Private Speech Trial		Second Private Speech Trial	
	M	*SD*	*M*	*SD*
Amount of Private Speech (utterance/min)[14]	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).*

[14] 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.

The results of this model, shown in Table 2.2, reveal three main findings. First, there was a main effect of amount of PS on performance (β = -0.15, 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 baseline competency was insignificant (p = 0.143), 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 = 0.44, 95% CI = [0.31, 0.57], 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)[15].

[15] 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 = -0.13, 95% CI = [-0.24, -0.02], p = 0.017) and no significant interaction (p = 0.253).

Table 2.2: The results of a Type III Multilevel Model for Testing the Effects of Private Speech on Performance and an Expertise Reversal

	Performance in the Private Speech Condition (number of turns)		
Predictors	*β*	*std.95% CI*	*p*
(intercept)	0.01	-0.13 – 0.14	0.886
Baseline Competency	0.44	0.31 – 0.57	<0.001
Amount of PS	-0.15	-0.25 – -0.05	0.002
Baseline Competency * Amount of PS	-0.09	-0.21 – 0.03	0.143
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		

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 2.3 (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 2.1, 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 2.3: 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 Expression	Expressing optimism, encouragement, and praise	"Good job", and "that's getting 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.	

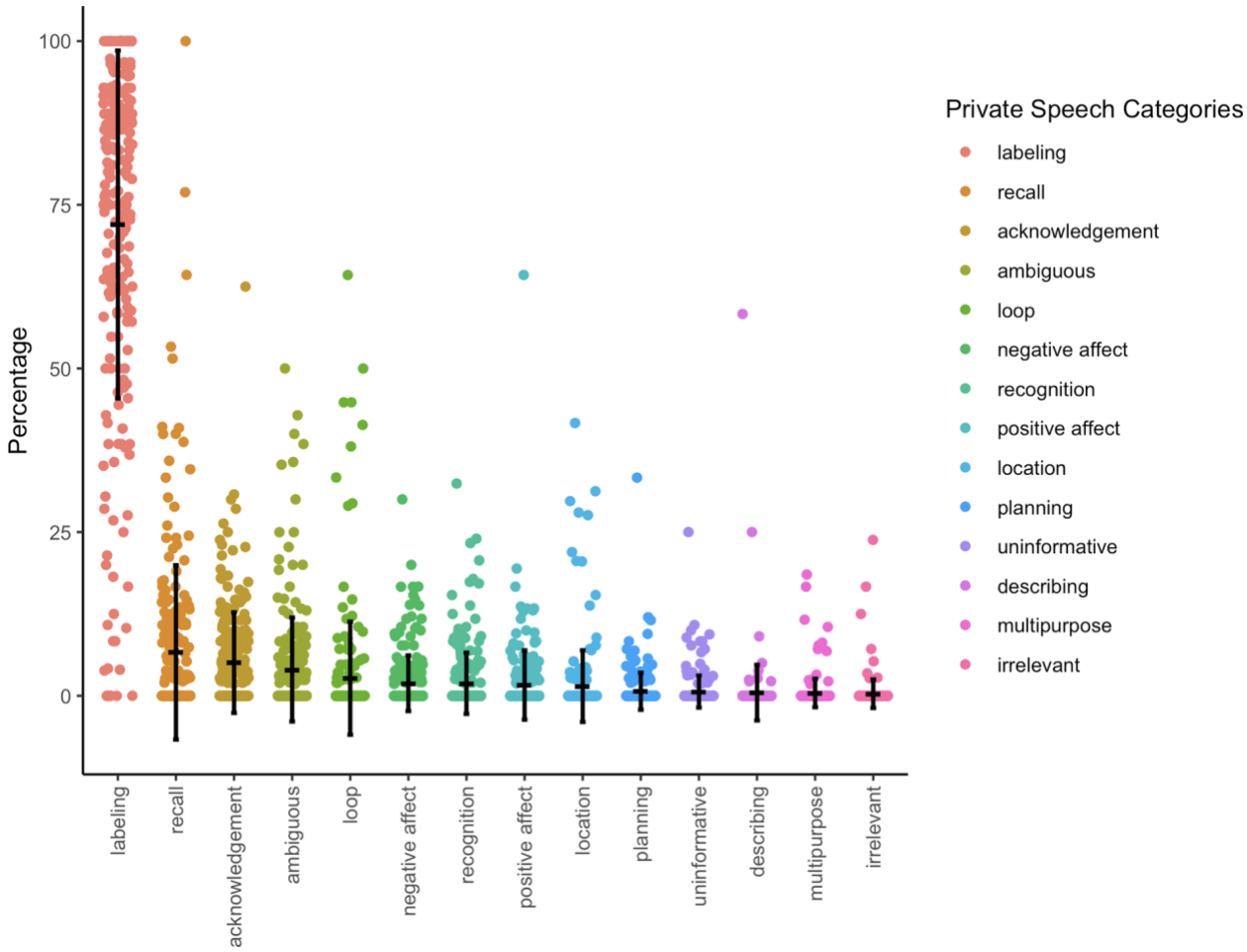

Figure 2.1: Distribution of Private Speech Content

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[16]. 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

[16] 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.

improved performance. With respect to what might underlie the beneficial effects of private 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[17]. 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 individual, and not help (or even hurt) if the task is relatively easy for an individual. In addition,

[17] Although we did not find any systematic order effect between the two Private Speech trials (see Table 2.1 of Results), an order effect could nonetheless exist between the first two (Baseline) trials and the next two (Private Speech) trials.

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 Appendix B, 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.

Acknowledgments

Chapter 2, in full, is a reprint of the material as it appears in Consciousness and Cognition 2023. Xinqi, Guo; Dobkins, Karen, Elsevier. The dissertation author was the primary investigator and author of this paper.

Appendix B

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 2.4 and Supplementary Table 2.5 show the descriptive statistics of experiential questions related to specific private speech trials and variables that assess participants’ overall experience, respectively.

Table 2.4: 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)

Table 2.5: 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 (β = 0.10, p = 0.039, 95%CI = [0.00, 0.19], see Table 2.6). 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*).

Table 2.6: Association between (Level 1) Objective and Subjective Extent of Private Speech.

	Extent of PS Usage		
Predictors	*β*	*std.95% CI*	*p*
(intercept)	-0.01	-0.17 – 0.15	< 0.001
Objective amount of PS	0.10	0.00 – 0.19	0.039
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 minut

CHAPTER 3 PRIVATE SPEECH IMPROVES COGNITIVE PERFORMANCE IN YOUNG ADULTS

Abstract

The current study investigated the relationship between private speech usage and cognitive performance in young adults. Participants (n = 106, mean age = 20.14 years) were instructed to complete a visual-spatial working memory task while talking out loud to themselves as much as possible (Private Speech condition). We found that participants performed better on trials for which they produced a greater amount of private speech. To establish causality, we further found that participants performed better in the Private Speech condition than in a condition in which they were instructed to remain silent (Quiet condition). These beneficial effects of private speech were not moderated by task difficulty, which was manipulated by varying image labelability. However, participants who used more private speech during the task, as well as those who reported greater use of self-management private speech in everyday life, showed the greatest benefits. These findings have implications for real-world educational/instructional settings.

Introduction

Humans possess the unique ability to talk to themselves, and this self-talk can be in the form of "inner speech" (i.e., thinking inside the head) or "private speech" (i.e., talking out loud to oneself). Given the general pervasiveness of self-talk, a natural question arises as to whether it has beneficial effects on everyday functioning. In the laboratory, this question has been studied mostly in the domain of *cognitive* functioning, with different methodological approaches used to study inner vs. private speech. To investigate the benefits of *inner* speech, "articulatory suppression" studies compare cognitive performance between conditions in which inner speech is vs. is not diminished/suppressed, with results demonstrating that inner speech facilitates performance on (some) cognitive tasks (see Nedergaard et al., 2022 for a review). To investigate the benefits of *private* speech, studies compare performance between conditions in which participants are instructed to use private speech vs. those in which they are either given no instruction (and presumably do not talk out loud) or are explicitly instructed to not talk out loud[18]. To date, the vast majority of studies demonstrating the benefits of private speech have been restricted to *children* (Fernyhough & Fradley, 2005; Lee, 1999; Winsler et al., 2007), likely owing to the fact that private speech is known to be prominent in children, but not adults (see Berk, 1986; Winsler et al., 2003, for empirical evidence and Vygotsky, 1987 for theory). Still, adults do use private speech in their everyday lives, with studies showing the highest frequencies of private speech usage 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 &

[18] 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.

Tarulli, 2009). Because these previous adult studies were designed to measure *spontaneous* private speech under different task conditions, rather than *instructing* adults to talk out loud, the question of whether private speech benefits cognitive performance in adults is still open. As such, the current study was designed to fill the gap in the literature, directly testing whether instructing adults to use private speech improves their cognitive performance.

The current study served as a follow-up to our previous study (Guo & Dobkins, 2023), in which we used a within-person correlational design to ask whether adults' amount of private speech while performing a cognitive (visual-spatial working memory) task is positively associated with their performance on that task. The task consisted of a card-matching game, called "Concentration Cat" (iOS App), wherein players had to find hidden pairs of matching images within an array by tapping/revealing two cards at a time. For each participant, performance was measured in two "Private Speech" trials, in which they were instructed to "finish the game in as few turns as possible, while talking out loud to themselves as much as possible" (and their amount of private speech, measured in utterances/minute, was determined for each trial). The results of this study showed that participants performed significantly better on trials for which they produced a greater amount of private speech. Because the vast majority of the content of their private speech was found to be *strategic* in nature (for example, using words related to what and where the hidden images might be), rather than in *response* to performance (for example, using words with positive affect after having found a hidden image), we argued that our correlational findings were likely to reflect a causal relationship, whereby the use of private speech benefits performance. Still, because that study was correlational in nature, the results could not provide conclusive evidence that private speech benefits performance. Therefore, the current study aimed to establish causality using an *experimental* approach,

comparing performance between conditions where participants are *instructed* to talk out loud vs. *not* talk out loud, counterbalancing the order of the two conditions across participants.

Note that in our previous study, we did have a condition (called the "Baseline" condition) in which we measured performance on the card-matching game when participants were *not* instructed to (and rarely ever did they) talk out loud, which, in theory, could have been compared to the two Private Speech trials. However, this was not possible, as we intentionally designed the study with the Baseline condition first as we were concerned that, if we randomized the order of the Baseline vs. Private Speech 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)[19]. The purpose of the Baseline measure was to serve as a trait measure of *competency* on the task so that we could ask an additional question; was the observed relationship between amount of private speech and performance stronger for participants with lower baseline competency, a notion that is in line with an "Expertise Reversal Effect" (observed in domains such as sports performance and second language learning, see Guo & Dobkins for discussion). Although we found no evidence for an expertise reversal effect, this null result could have been driven by having too narrow a range of baseline competency across our cohort of participants (ages 18 to 33 years, $M = 20.13$, $SD = 1.91$), an issue we return to below.

In sum, the main goal of the current study aimed was to directly test the impact of private speech on cognitive performance. We used the same card-matching game as in our previous

[19] 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 underestimation, and a "practice effect" will result in an overestimation, of this beneficial effect.

study (Guo & Dobkins 2023), but here, conducted a *Manipulation* analysis, comparing performance between two conditions, i.e., Private Speech and Quiet, counterbalanced in order across participants. As in our previous study, we presented two back-to-back trials of the Private Speech condition in which participants were instructed to "talk out loud as much as possible" during the game. And, we presented two back-to-back trials of the "Quiet" condition in which participants were explicitly instructed to "be quiet" during the game, which differed from our previous Baseline condition that provided no instructions regarding talking out loud. Collecting *two* trails per condition (as opposed to just one) had the additional benefit of allowing us a direct replication of our previous *Correlational* analyses, which showed that individuals performed significantly better on Private Speech trials for which they produced a greater amount of private speech (see *above*).

A second goal of the current study was to investigate whether the effects of private speech (in both our *Manipulation* and *Correlational* analyses) vary as a function of task difficulty. This question was inspired by studies showing that adults use spontaneous private speech more so under challenging situations (see *above*) as well as evidence from the child private speech literature showing that the frequency of private speech varies with task difficulty (Fernyhough & Fradley, 2005). As we have pointed out previously (Guo & Dobkins, 2023), the relevant dimension that may affect the degree of benefit of using private speech is the *relationship* between task difficulty and participant competency/expertise, noting that the former can be manipulated by varying the stimuli (e.g., Fernyhough & Fradley, 2005), whereas the latter can be varied by testing participants with different competencies/expertise (e.g., "verbalization" studies in children vs. adults, Kray et al., 2008). In our previous study (Guo & Dobkins, 2023), we used a single version of the card-matching game (in which all the images were easy to label)

with the hope that its "perceived difficulty" would vary sufficiently across participants to see the effect of difficulty. Here the assumption is that baseline competency (determined from the condition in which participants did not talk out loud) can be considered a proxy for how difficult one finds the card-matching game (i.e., low competency reflects greater difficulty). Our study found no moderating effects of baseline competency/perceived difficulty on the relationship between amount of private speech and performance, although this might have resulted from there not being enough variation across participants in the perceived difficulty of the task to show an effect.

To more directly test the potential moderating effects of task difficulty, the current study tested participants in *two* versions of the card-matching task, designed to create variation in task difficulty. To achieve this, we manipulated the labelability of the images in the card-matching game, using both easy-to-label images, as in our previous study, as well as hard-to-label images that are more abstract in nature. The images we used were Tangram images, which were piloted in a separate set of participants to determine ones that were easy- vs. hard-to-label, based on reaction times (see *Methods*). The notion that using Easy vs. Hard images might produce differences in task difficulty comes from previous literature showing that visual working memory performance tends to be better when images are easy to label or meaningful, as opposed to hard-to-label or abstract (e.g., Brady et al., 2016; Brady & Störmer, 2021; Souza et al., 2021; Souza & Skóra, 2017; Sobrinho and Souza, 2023). For instance, Brady et al. (2016) showed that visual working memory was significantly better for real-world objects (which were easy to label) vs. colors (many of which were non-primary colors and were hard to label). Because they used an articulatory suppression paradigm to suppress inner speech, their findings suggest that the labelability effect is not the result of verbal-encoding differences, and that it therefore must occur

at a pre-cognition level. The exact mechanism underlying these effects is not fully understood, but converging evidence suggests that meaningful images (which are easy-to-label), more so than scrambled or unrecognized images (which are hard-to-label), activate prior knowledge in long-term memory, which serves to expand the capacity of working memory (Brady & Störmer, 2021; Asp, Störmer & Brady, 2021). Investigating the moderating effects of task difficulty was a confirmatory hypothesis in our pre-registration.

A third goal of the current study was to investigate whether the effects of private speech (in both our *Manipulation* and *Correlational* analyses) vary as a function of an individual's natural tendency to use private speech in their everyday life. Specifically, we wondered if the benefits of talking out loud, when instructed to do so (in an experimental setting), are greatest for people who have a natural "fluency" in talking out loud to themselves. In the opposite vein, for people who *never* talk out loud to themselves, asking them to do so might *impair* their performance. To test this, the current study employed the Self-Talk Scale (Brinthaupt et al., 2009), which asks participants to self-report their self-talk usage in everyday life. Although this scale does not query specifically about whether one's self-talk entails inner vs. private speech, we modified the scale to our purposes asking only about private speech usage. In previous studies, it has been shown that responses on this scale do, in fact, predict performance. For example, Shi et al. (2017) showed that participants who report using the self-management type of self-talk more frequently also performed better when assigned to give a persuasive public speech. Likewise, the current study predicted that habitual use of certain types of private speech (e.g., self-management) might moderate the benefits of private speech on card-matching game performance.

Method

The hypothesis, study design, exclusion criteria, and analysis plan were preregistered: https://osf.io/uz9kf

Participants

Participants were undergraduate students recruited through a participant pool run by the Department of Psychology at the University of California San Diego between November 2022 and January 2023. 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 credits. The study was approved by the Institutional Review Board at our university.

The collected sample consisted of 110 participants, a sample size that was determined by a priori power simulation (see pre-registration). Four participants in total were excluded for the following reasons: not following the proper procedure (n = 1), failing an effort check[20] (n = 1), or because their private speech was not recorded due to experimenter error (n = 2). The demographics of the remaining 106 participants whose data were analyzed were as follows: Age ranged from 18 to 33 years (M = 20.14, SD = 2.12), the gender identities were 69.6% women, 25.5% men, 3.0% non-binary, and 1% "prefer not to say", and the ethnicities were 51.0% Asian, 11.8% White, 20.6% Hispanic, 3.9% Middle Eastern or North African, 2.0% Black/African American, 8.8% mixed, and 2.0% "prefer not to say".

General Study Design

In this study, there were two main manipulations. The *first* manipulation was “Speech Condition”, with trials being either Quiet (participants were asked to *not* talk out loud while

[20] At the end of the study, we surveyed participants over Qualtrics about their level of effort/engagement during the study, with two separate questions. We considered a participant as “failing” the effort check if they chose “I did not follow the instructions of the game and instead tried to finish the game as quickly as possible.” in the first question and “I did not read those questions carefully, and instead, I answered them as quickly as possible.” In the second question.

performing a card-matching task) or Private Speech (participants were asked to talk out loud as much as possible during that task). The *second* manipulation was "Labelability", with the images being either easy-to-label ("Easy" condition) or hard-to-label ("Hard" condition). This 2 × 2 design resulted in four total conditions, and participants were tested with two trials per condition, resulting in eight total trials per participant. The data from these eight trials allowed us to conduct two different types of analyses, separately for the Easy and Hard condition. The *Correlational* analyses served as a replication of Guo and Dobkins (2023), in which we showed that participants performed significantly better on the Private Speech trial for which they produced a greater amount of private speech. The *Manipulation* analyses were conducted to establish causality, comparing performance between the Quiet and Private Speech conditions whose order was counterbalanced across participants.

Apparatus and Materials

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 × 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 × 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 eight trials, and thus we needed 96 unique images (i.e., 12 per trial).

Creating Stimuli for the Easy and Hard Labelability Conditions

In our original study (Guo & Dobkins, 2023), we employed images that were easy-to-label. In the current study, our goal was to replicate this previous study using easy-to-label images, and in addition, add a condition in which the images were hard-to-label, as a way of varying the *difficulty* of the task (see *Introduction)*. Rather than use our old set of easy-to-label images, we decided to create a new bank of images, wherein both the Easy and Hard images share the same low-level visual features. To this end, we used Tangram images (from an online source, Tangram Channel, 2015), each of which consists of seven geometric pieces, including triangles and a square that can be rearranged to form various figures and shapes. We began by collecting a large pool of candidate stimuli that consisted of 245 unique black tangram patterns with white backgrounds. Some of the patterns resemble familiar concepts (like digits, letters, and real-world objects), which are likely easy to label, whereas others consist of unfamiliar shapes, which are likely harder to label. Capitalizing on this variety, our goal was to obtain a set of Easy vs. Hard-to-label images, based on how long it took participants to label the images. To this end, we tested a separate set of 13 participants (undergrads at our university), who were tasked with *labeling out loud* each image as quickly as possible, with the idea that short and long response times were indicative of easy- and hard-to-label images, respectively. Participants were tested on a laptop in the lab (using PsychoPy, v2022.1.3, Peirce, 2007), which presented each tangram image on the screen, one at a time (in a randomized order). Participants were instructed to press a

spacebar as soon as they finished labeling each tangram out loud. Each participant began with a practice session of five images (not contained in the set of 245), with the experimenter present to make sure the participant was following the instructions. The experimenter then stepped out of the testing room, so as to not make the participant uncomfortable while doing the task on the 245 images.

To select 48 images for the Easy and Hard conditions, we ranked the tangram images based on the average time (across participants) taken to hit the space bar (which is the operationalization of how long it took to label the images out loud). The 48 patterns with the shortest and longest time to respond were assigned to be used in the Easy vs. Hard conditions, respectively (Hard: M = 4.63 secs, SD = 0.45; Easy = 1.90 secs, SD = 0.20; Cohen's d = 7.28). The 96 images were inspected by the first author to ensure that any two were highly unlikely to receive the same label (i.e., that all images were categorically and visually distinct). The 48 images for the Easy and Hard conditions were randomly divided into four sets of 12, each of which constitutes the stimuli of one trial. The assignment of images to trials as well as the locations of images in the card-matching game were kept the same across participants. See Figure 3.1 for example Tangram images.

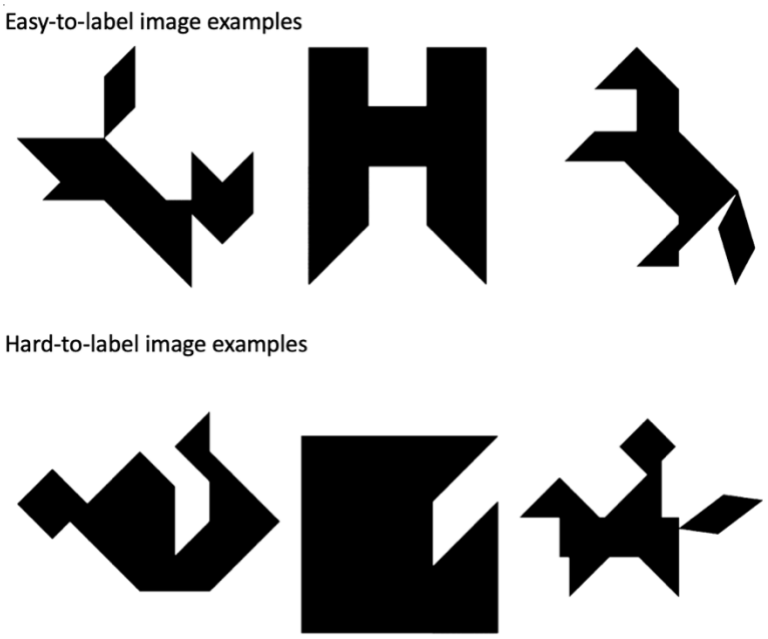

Figure 3.1. Example Tangram Images

Trait-level Private Speech Questionnaires

We obtained a *trait*-like measure of how frequently participants used private speech in their everyday life by administering the Self-Talk Scale ("STS", Brinthaupt et al., 2009), which is a 16-item self-report questionnaire that assesses the individual tendency and reasons for talking to oneself[21]. The STS was developed, and has been validated, in undergraduate student populations, and assesses self-talk in general (either private speech or inner speech). Because the current study investigated private (and not inner) speech, we modified the STS so that the

[21] Alternative questionnaires assessing self-talk behaviors exist, but they are comparatively less suitable for the specific aims of our study. For instance, the Self-Verbalization Questionnaire ("SVQ", Duncan & Cheyne, 1999) assesses the frequency of private speech at a trait-level. However, SVQ primarily focuses on assessing attentional and mnemonic functions of self-talk, while other functions that self-talk may serve are largely ignored. On the other hand, our modified STS surveys a broader range of situations when private speech might be used and is a better fit for assessing individual differences in trait tendency of private speech. Additionally, SVQ was developed a long time ago, as a result, some of its items have become outdated. For example, the item "I sometimes think out loud to myself when I'm looking for a number in the phone book" might perplex our young adult participants who are unfamiliar with phone books.

leading statement was specific to private speech, as follows: "I talk to myself *out loud* when..." followed by a situation of interest. Participants provided their responses using a 5-point Likert scale with options ranging from "Never" to "Very Often". The STS has four subscales (each subscale comprises four items) to capture different aspects of everyday usage of private speech:

1- Social Assessment measures the tendency to engage in private speech to replay conversations to oneself and envision the reactions of others. For example, “I talk to myself out loud when I’m imagining how other people respond to things I’ve said”.

2- Self-Reinforcement measures the tendency to engage in private speech when experiencing a sense of accomplishment or when a positive event has occurred. For example, “I talk to myself out loud when I want to reinforce myself for doing well”.

3- Self-Criticism measures the tendency to engage in private speech when criticizing oneself for things said or done and showing discouragement. For example, “I talk to myself out loud when I’m really upset with myself”.

4- Self-Management measures the tendency to engage in private speech when self-directing and deciding on the appropriate actions or words to say. For example, “I talk to myself out loud when I’m mentally exploring a possible course of action”.

The STS has been used to ask whether certain types of self-talk predict trait characteristics. For example, Brinthaupt et al. (2009) reported that (higher) responses on the Self-Criticism sub-scale predict (lower) self-esteem.

Participants also filled out some in-house measures of frequency, comfort, and attitudes about private speech, as follows:

Frequency: "In general, how often do you talk out loud to yourself?" Participants rated their frequency of talking out loud on a 10-point Likert scale, ranging from "1 - Never" to "10 - Very Often". In the current study (see *Results)*, we conducted exploratory analyses asking whether this Frequency score moderated any of the observed effects (in both the *Correlational* and *Manipulation* analyses).

Comfort: "In general, how comfortable are you talking out loud to yourself?" Participants rated their comfort level on a 10-point Likert scale, ranging from "1 - Not at all" to "10 - Completely/Entirely."

Tendency: Because we found a high correlation between "frequency" and "comfort", $r(77^{22}) = 0.68$, $p < .001$, in line with our pre-registration, we created an average score of the two, which we refer to as "Tendency".

Attitudes: "Do you think talking to oneself out loud has a negative societal taboo attached to it?" using a 10-point Likert scale, with "1 - Not at all" and "10 - Very much" as the two extremes. The mean score on this metric (normalized by subtracting by 1 and dividing by 9, so that 1.0 was the max and 0 was the minimum) was $M = 0.48$ ($SD = 0.26$), indicating that participants felt, on average, that talking out loud to themselves was moderately a taboo. We did not use this metric in any of our analyses, but include it only to present a descriptive statistic regarding participants' attitudes about talking out loud to oneself.

[22] The degrees of freedom is lower (than 106) for this correlation analysis since we did not have the comfort question in the questionnaire for the first 30 participants due to error.

In the current study, we explored whether scores on any of the four subscales of the STS or the 1 in-house Tendency construct moderated any of the observed effects (in either the *Correlational* and *Manipulation* analyses). This meant that we tested for potential effects of five different "*Trait-Private Speech (PS)"* metrics.

In-lab Procedure

Upon arrival at the lab, participants were asked to complete the Trait-PS Questionnaires (filled online over Qualtrics) in a waiting area of the lab. They were then guided into the testing room, where they were informed that they would be playing a card-matching game, which was explained to them through a pre-recorded video demonstration on a laptop computer. The video demonstration featured a 2 × 3 array of face-down cards with patterns different from those used in the actual trials. During the video, the experimenter paused periodically to elaborate on the rules and goals of the game. Next, the experimenter proceeded by setting up participants to play *eight* actual trials of the game on an iPad, through an iOS app called "Concentration Cat", which was specifically developed for our study by a professional. The experimenter stepped outside the testing room during all eight trials, so as to not make the participant uncomfortable, and only came back in between the trials to deliver instruction for the next trial. The eight trials were presented in a pre-designed order, as follows.

Labelability of the images (Easy vs. Hard) was blocked and counterbalanced across participants, with half of the participants starting with four Easy trials (which we designate as *Block* Order = 0), and the other half starting with four Hard trials (which we designate as *Block* Order = 1). Within each block of four trials, the order of the Speech condition (Quiet vs. Private Speech, described further below) was counterbalanced across participants, with half of the

participants starting with the two Quiet trials first (which we designate as *Trial* Order = 0), and the other half of the participants starting with two Private Speech trials first (which we designate as *Trial* Order = 1), and this Trial Order was maintained across the two labelability blocks. This resulted in participants being counterbalanced into four different orders of the eight trials, as shown in Figure 3.2.

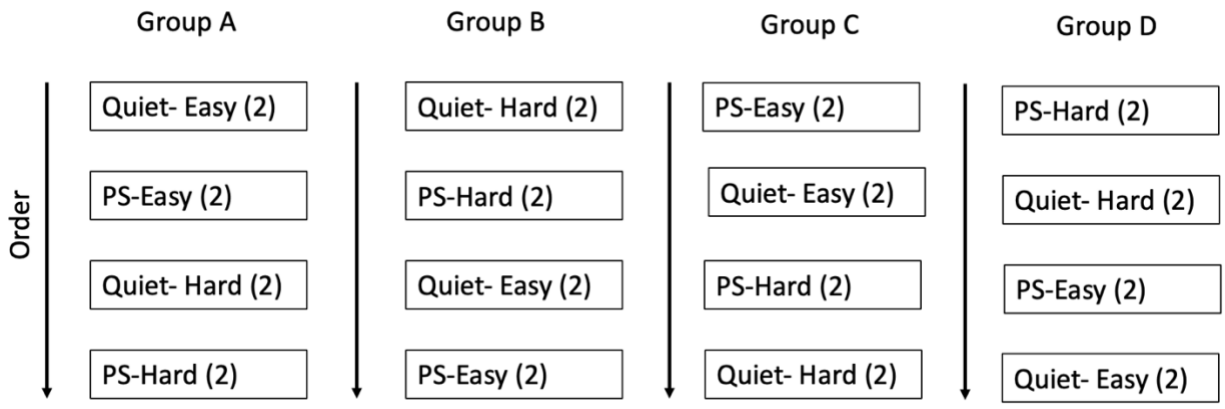

Figure 3.2. Counterbalanced Trial Sequence for Labelability and Speech Conditions

On Quiet trials, participants were asked to finish the game in as few turns as possible, and to not talk out loud. Specifically, they were told "*Please finish the game quietly and use as few taps as you can. You will see the time and taps you used after each trial. But the only goal is to use as few taps to finish the game, and we do not care about the time taken to finish the game in this study. You can finish the game at your own pace. I (the experimenter) will be outside, and the door will be closed."*. Note that rarely ever did a participant spontaneously talk out loud in this condition (see *Results*). On the Private Speech trials, participants were given similar instructions about the task itself but were asked to "talk out loud as much as possible" during the game. Specifically, they were told: "*Please finish the game using as few taps as you can. You*

will see the time and taps you used after each trial. But the only goal is to use as few taps to finish the game, and we do not care about the time taken to finish the game in this study. You can finish the game at your own pace. Talk to yourself audibly or externally throughout the game or as much as you can. You can use the 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 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 the amount and content of their private speech (see *below*). Also unbeknownst to the participants, we used a screen capture function on the iPad to collect two pieces of information: (1) number of turns, and (2) time to complete the trial. Both (1) and (2) were automatically provided by the iOS App after each trial. (1) was used as our main performance measure, and (2) was used to compute rate of private speech. The screen and audio recordings were collected for all eight trials (the Baseline and Private Speech trials). After finishing all eight trials, the experimenter came back to the testing room again with a laptop, where the Qualtrics questionnaire was loaded, and instructed the participant to answer two questions about their effort level in the experiment (see "failed effort check", *above*) as well as demographic questions. In an email after the experiment, the participants were debriefed about being secretly audio-recorded during the experiment, and they were given an audio consent form to indicate if they agree for their audio to be analyzed anonymously for research purposes.

Measures

Performance Measurement.

As in our previous study (Guo & Dobkins, 2023), the main measurement of performance for each trial was the "*number of turns"* (i.e., pairs 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). The "number of turns" was z-scored (within the relevant grouping[23]) and converted into its additive inverse, so that higher numbers represent better performance, which facilitates understanding of figures depicting performance as a function of other variables. For the *Correlational* analyses, the two Quiet trials were averaged and used as a measure of *baseline competency* on the task, to explore the Expertise Reversal Effect (see Guo & Dobkins, 2023).

Amount of Private Speech (PS).

The current study used utterances/minute as the objective measure of amount of private speech. The choice of this metric (as opposed to *total* utterances or time) is justified in our previous study (Guo & Dobkins, 2023), noting that it has also been used in previous private speech studies (Duncan & Cheyne, 2001; Kronk, 1994; Fernyhough & Fradley, 2005; Mulvihill et al., 2021). As a first step, the audio recordings of participants' private speech were analyzed offline by the first author and her research assistants. For private speech trials in English (86.5% of all trials), an automatic speech recognition tool named Whisper (Radford et al., 2022) was

[23] For the *Correlational* analyses, this meant that the transformation was done separately for Quiet trials (to obtain a measure of baseline competency) vs. the two Private Speech trials. For the *Manipulation* analyses that included both the Easy and Hard condition in the same model, this meant that the transformation was done using data from all eight trials). For the *Manipulation* analyses that included only one of the two labelability conditions, this meant that the transformation was done separately on the four Easy vs. the four Hard trials.

used to generate the initial transcription. The Whisper-transcribed utterances were then reviewed and edited by the first author to ensure accuracy. The initial transcriptions of non-English languages were performed by the first author or her research assistants who know the language (Mandarin: 9.8% of trials, 1.0% Korean, 1.0% Turkish, 1.7% 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.

Next, data were entered into a spreadsheet in units of "Utterances", defined as an audible verbal unit separated by differences in semantic meaning or at least one second of temporal distance (Frausel et al., 2020; Rowe, 2012; Rowe & Goldin-Meadow, 2009). For example, "Dog at the top right corner" would be considered as one utterance, whereas "Is the dog here? Nope." would be considered as two utterances. We tested inter-rater reliability for a random subset of 10 participants (40 trials) by having a second transcriber, in addition to the first author. The data from these 10 participants showed very high inter-rater reliability in quantifying the number of utterances (Easy: ICC = 0.962; Hard: ICC = 0.956). Importantly, the transcribers were blind to whether the trials were from the Easy or Hard condition when transcribing. As a final step, utterances/minute was calculated as the number of utterances divided by the time to finish the trial.

Note that for our *Correlational* analyses, which used a within-person approach that replicates Guo & Dobkins (2023), we needed to transform amount of private speech. As we explain in that previous study, 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. 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 3.2% 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). For the *Manipulation* analyses, where we asked whether amount of PS moderates any of the observed effects, we z-scored the average amount of private speech across the two Private Speech trials (separately for Easy vs. Hard), and used this as a Level 2 variable.

Private Speech Content Coding

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. As in our previous study (Guo & Dobkins, 2023), we placed each utterance into one of 14 categories, outlined in Table 3.1 (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. We tested the inter-rater reliability of content coding with the same second transcriber and the same 10 participants, who were involved in quantifying inter-rater reliability for coding utterances (*above*). The data from these 10 participants demonstrated high inter-rater reliability in coding private speech content (κ = .633, percentage agreement [categorization of content being the

same] = 73.5%, 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, 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.

Table 3.1: Private Speech Content Categories, Definitions, and Examples from the Current Dataset

Category	Definition	Examples from the study
Labeling	Labeling the card patterns or showing an attempt to label.	“Dog”, and “apple”
Ambiguous or unclear	Audible but unintelligible whispering	The content cannot be coded, but the quantity was estimated.
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”
Describing	Verbally describing stimuli, but no label	"A yellow and round… thing"
Positive Affect Expression	Expressing optimism, encouragement, and praise	"Good job", and "that's getting better"
Negative Affect Expression	Expressing pessimism, discouragement, and criticism.	"Looks like I messed up already", "oh man!", and profanities.
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"
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”
Planning	Planning for actions. Self-guided, self-managing attempts.	"Ok start from the top right", "do not tap this", “going to try the edges"
Acknowledgment	Spontaneous reactions that are not emotional expressions.	"Ah", "ha", "I don't know what that is", "what?", "alright", and "ok"
Uninformative	Not serving any specific function other than showing the individual is paying attention to the game.	"dudududu", "this one", and "let's see"
Rehearsing	Rehearsing the previously seen stimuli when revealing new cards.	"Cat, bathtub, key, dog, blanket”
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.	

Analyzing the Effects of Block and Trial Order

Although Block and Trial Order were counterbalanced across participants, so that they should not *account* for our findings, we explored their influence on performance nonetheless. Note that while *Block* Order is simple to conceptualize (referring to whether a participant was tested first with a block of four Easy trials (Block Order = 0) vs. four Hard trials (Block Order = 1), *Trail* Order is a bit more complicated. This is because Trial Order (referring to whether a participant was tested first with two back-to-back Quiet trials (Trial Order = 0) vs. two back-to-back Private Speech trials (Trail Order = 1) is a concept that exists *within*, but not across, Blocks. As this was not part of our pre-registration and because the effects are not germane to the focus of this study, we present the effects of Block and Trail order in Appendix C.

Data Exclusion

At the *participant* level, we applied separate exclusion criteria for the *Correlation* and *Manipulation* analyses since participants who were disqualified for one analysis could qualify for the other analysis, and we wanted to retain as much data as possible for each analysis. As described above, for the *Correlational* analyses, the average performance across the two Quiet trials was used as a measure of participants' baseline competency, separately for the Easy vs. Hard labelability conditions. As in our previous study (Guo & Dobkins, 2023), a participant was excluded from these analyses if their baseline competency was three standard deviations worse than the average across participants, computed separately for the Easy vs. Hard conditions. This exclusion criterion resulted in two participants excluded from the Easy condition (leaving 104) and three participants excluded from the Hard condition (leaving 103). For the *Manipulation* analyses, the two Quiet trials were treated as individual trials rather than being averaged to

calculate a baseline competency measure. Thus, no participant was excluded from the *Manipulation* analysis, and data from all 106 participants were retained for analysis.

At the *trial* level, exclusion criteria were applied separately for each of the four conditions (i.e., Quiet-Easy, PS-Easy, Quiet-Hard, PS-Hard). A trial was excluded if performance on that trial was three standard deviations worse than the trial-wise average performance of the condition. Accordingly, four trials were excluded for Quiet-Easy, four trials were excluded for PS-Easy, four trials were excluded for Quiet-Hard, and four trials were excluded for PS-Hard. 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).

Results

Descriptive Analyses

Descriptive data of means and distributions of study variables are presented from 832 trials (8 trials × 106 participants after excluding sixteen outlier trials, see *Methods)*[24]. With regard to *performance,* the mean number of turns for the four different conditions were as follows: Quiet-Easy trials = 28.18 (*SD* = 6.26), Quiet-Hard trials = 33.04 (*SD* = 7.76), PS-Easy trials = 26.55 (*SD* = 5.70), PS-Hard trials = 30.81 (*SD* = 7.69). Formal statistical analyses that compare across the four conditions are presented in the *Manipulation* analyses section (below), but we note up front that participants performed worse when images were hard-to-label, which confirms that this condition was, in fact, more difficult, as we intended it to be.

[24] Note that there is some interdependence in these means as they are created from two trials per participant. Still, they provide a reasonable estimate of the means. The actual model-estimated means, which accounts for this interdependency, are presented later in the Results, in Figure 3.3.

With regard to *Amount of Private Speech (PS)*, we had hoped that the Quiet condition would yield no private speech, since participants were explicitly told not to talk out loud. However, 3.0% of the Quiet trials (Easy: 2.1%, Hard: 0.9%) nonetheless contained some spontaneous utterances, and we chose to retain these rare trials in our analysis[25]. For the Private Speech trials, the mean number of utterances/minute for the Easy condition was 22.23 (*SD* = 10.16), which was similar to the values observed in our previous study that also used easy-to-label images (i.e., *M* = 27.56, *SD* = 11.26). For the Hard condition, the mean number of utterances/minute was 14.76 (*SD* = 8.59) was significantly lower than observed in the Easy condition ($p < 0.001$[26]).

With regard to private speech *content*, we present the distributions for the Easy and Hard conditions in Table 3.2. For the Easy condition, the private speech content distribution was very similar to that observed in our previous study, $\chi^2(13) = 6.91$, $p = 0.907$. Although a chi-squared analysis revealed no difference in content distribution between the Easy and Hard conditions of the current study ($\chi^2(14) = 15.98$, $p = 0.314$), visual inspection reveals that the category "labeling" was much higher in the Easy condition (62.23%, *SD* = 30.41%) than the Hard condition (36.70%, *SD* = 31.21%), and conversely, that the category "describing" was much higher in the Hard condition (3.24%, *SD* = 8.71%) than the Easy condition (0.95%, *SD* = 3.71%). These differences are intuitive, in that participants moved from "labeling" to "describing" when the images were hard-to-label.

---

[25] We decided not to exclude these trials because they were rare and because they had low amounts of private speech. On average, the average amount of private speech in these trials were in the 9th percentile for Easy and 8th percentile for Hard. Importantly, the direction and level of significance of the results remained unchanged in an exploratory analysis in which these trials were removed.

[26] To analyze differences in amount of private speech between the two labelability conditions, we had to use a multilevel model, since each participant provided two private speech trials for each labelability condition. In this model, the dependent variable was amount of PS, and the predictor term was labelability condition (Easy vs. Hard), with participant included as a random intercept effect. This analysis revealed that the amount of PS was significantly higher in the Easy vs. the Hard condition ($\beta = 0.72$, 95%CI =[0.60, 0.82], $p < 0.001$).

Table 3.2: Private Speech Content Distribution as A Function of Labelability

Category	**Easy (%)**	**Hard (%)**
Labeling	62.20	36.70
Acknowledgement	9.94	17.20
Uninformative	4.36	9.45
Recall	4.30	6.86
Rehearsing	4.03	2.19
Ambiguous	3.51	7.32
Positive	2.29	3.97
Negative	2.27	5.00
Recognition	1.91	2.67
Irrelevant	1.32	1.64
Planning	1.20	1.41
Location	1.18	1.67
Describing	0.95	3.24
Multipurpose	0.40	0.36
Other	0.10	0.36

Note. The rows are sorted based on the proportions of private speech categories for the Easy condition, with the top row representing the category with the largest proportion within the Easy condition, and the bottom row representing the category with the smallest proportion. However, it is apparent from the values for the Hard condition that this order was different, as some lower rows have larger proportions than the upper rows.

Consistent with Guo and Dobkins (2023), “Labeling” (e.g., “dog”, “cow”) was the most frequent category for both the Easy and Hard condition, which appears to be a type of strategizing to remember the identity of the cards (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", "rehearsing", and "describing") was 74.74% and 53.31% for the Easy and Hard conditions, respectively. 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 low (Easy: 4.43%, Hard: 8.72%), with roughly half being positive, and half being negative, responses. As we have argued previously, the finding that the majority of utterances likely occurred *prior* to (as opposed to in response to) a correct/incorrect match is in line with the degree of private speech usage affecting performance rather than one's level of performance affecting their degree of private speech (see Guo & Dobkins, 2023). The direct testing of causality is addressed in the *Manipulation* analyses, below.

Correlational Analyses: Testing the Relationship between Amount of Private Speech and Performance

These analyses served as a replication of Guo and Dobkins (2023), in which we showed that participants perform significantly better on trials for which they produced more private speech. This previous study was conducted using easy-to-label images, whereas the current study employed both easy-to-label images (Easy condition) and hard-to-label images (Hard condition). While the purpose of the current study's Easy condition was to provide a direct[27] replication of Guo and Dobkins (2023), the purpose of the Hard condition was to test whether the effect of private speech on cognitive performance differs when the task is made more difficult, which was manipulated by using hard-to-label images (see *Descriptive Analyses*, above).

[27] We consider this a direct replication even though the Guo & Dobkins (2023) study employed easy-to-label stimuli that were taken from colored clip-art images of real-life objects, whereas the current study used tangram images that we empirically showed were easy-to-label. We believe this one difference between the two studies is minute.

As a first step in our analyses, we person-mean centered the amount of PS for each of the two Private Speech trials, which allowed us to look at Level 1 effects (see *Data Transformation* in *Methods*). The two Quiet trials were not person-mean centered and instead were averaged to obtain a measure of baseline competency (see *Methods*). Using a Type III sum of squares multilevel regression model, separately for the Easy vs. Hard condition, the dependent variable was performance on the private speech trials and the predictor terms were: 1) Level 1 Amount of PS (entered as a fixed effect), 2) baseline competency (entered as a fixed effect), with Participant included as a random intercept effect. As we did in our previous study, we also added an interaction term between (1) and (2), to check for an Expertise Reversal Effect, whereby participants who have lower baseline competency might show a stronger relationship between *amount of PS* and performance.

The results of our analyses are shown in Table 3.3 (*Left* panel: Easy, *Right* panel: Hard). Since the direction and effect size of the predictors' coefficients were largely consistent between the Easy and Hard conditions, we present a single narrative for both as follows. Replicating Guo and Dobkins (2023), these analyses revealed three main findings. *First*, there was a main effect of Level 1 Amount of PS on performance (Easy: $\beta = 0.13$[28], 95% CI = [0.04, 0.22], $p = 0.005$; Hard: $\beta = 0.17$, 95% CI = [0.07, 0.27], $p = 0.001$), with higher amounts of private speech being associated with better performance. In other words, participants performed better on the trial for which they produced a greater amount of private speech. *Second*, as might be expected, baseline competency predicted performance in the Private Speech condition, i.e., people who did better in the Quiet condition also did better in the Private Speech condition (Easy: $\beta = 0.47$, 95% CI = [0.33, 0.62], $p < 0.001$; Hard: $\beta = 0.40$, 95% CI = [0.25, 0.55], $p < 0.001$). *Third*, there was no

[28] The βs in the current and Guo and Dobkins (2023) are very similar.

significant interaction between Level 1 Amount of PS and baseline competency (Easy: p = 0.485; Hard: p = 0.171), meaning that the relationship between Level 1 Amount of PS and performance was invariant across participants with different levels of baseline competency. When we removed baseline competency and the interaction term from the model, the effect size of Level 1 Amount of PS remained identical, although the marginal R-squared of the model necessarily became smaller (Easy: 0.017; Hard: 0.035, data not shown).

To address whether the effect of Level 1 Amount of PS *differed* between the Easy and Hard conditions, we included both Easy and Hard trials into the same model. Here, the dependent variable was performance on the private speech trials and the predictor terms were: 1) Level 1 Amount of PS (entered as a fixed effect), 2) Labelability condition (entered as a fixed effect), and 3) the interaction between (1) and (2), with Participant included as a random intercept effect. Because we found no interaction between Level 1 Amount of PS and Labelability (p = 0.21), this suggests that the relationship between Level 1 Amount of PS and performance is comparable across the two levels of task difficulty.

Table 3.3: The Results of Type III Multilevel Models for Testing the Effects of Level 1 Amount of PS and Baseline Competency on Performance in the Easy (left) and Hard (right) conditions.

	Performance in the Private Speech Condition (additive inverse of the standardized number of turns)					
	Easy			Hard		
Predictors	*β*	*95% CI*	*p*	*β*	*95% CI*	*p*
(intercept)	0.00	-0.14 – 0.15	0.977	-0.01	-0.16 – 0.14	0.939
Level 1 Amount of PS	0.13	0.04 – 0.22	**0.005**	0.17	0.07 – 0.27	**0.001**
Baseline Competency	0.47	0.33 – 0.62	**<0.001**	0.40	0.25 – 0.55	**<0.001**
Level 1 Amount of PS × Baseline Competency	-0.03[29]	-0.11 – 0.05	0.485	-0.07	-0.17 – 0.03	0.171
Random Effects						
σ2	0.42			0.44		
τ00	0.36 $_{Participant}$			0.37 $_{Participant}$		
ICC	0.47			0.45		
N	104 $_{Participant}$			103 $_{Participant}$		
Observations	208			206		
Marginal R^2 / Conditional R^2	0.239 / 0.593			0.201 / 0.563		

Does Trait-PS Moderate the Effect of Level 1 Amount of PS on Performance?

Using Type III sum of squares multilevel regression models, separately for the Easy vs. Hard condition, we asked whether trait-level private speech usage (*Trait-PS*), obtained from the self-reports of participants, moderates the positive relationship between Level 1 Amount of PS

[29]If the regression coefficient of the interaction between the two predictors is negative, it means that the increase of one predictor will decrease the effect of the other predictor on performance. It is important to note that the interaction term is not significant here. However, if it were significant, then it would mean that the positive relationship between Level 1 amount of PS and performance gets weaker with increasing baseline competency.

and performance observed in the prior analysis. That is, we wondered whether the relationship between amount of PS and performance might be greater for people who report more usage of private speech in their everyday life. As explained in our pre-registration, because this analysis was exploratory, we tested separate models for each of the five different metrics of trait-PS (see *Methods*). For each of the five models, the dependent variable was performance and the predictor terms were: 1) Level 1 Amount of PS (entered as a fixed effect), 2) Trait-PS (entered as a fixed effect), and 3) the interaction between (1) and (2), with Participant included as a random intercept effect. The results of these analyses showed that for both the Easy and Hard conditions, there were no moderating effects of trait-PS for any of the five measures of trait-PS (Easy: all *p*s > 0.097, Hard: all *p*s > 0.134), meaning that the positive relationship between Level 1 Amount of PS and performance did not vary across participants with different levels of trait-PS. There was also no main effect of trait-PS on performance (Easy: all *p*s > 0.194. Hard: all *p*s > 0.207), meaning that participants who were high vs. low in trait-PS performed equally well on the private speech trials. Finally, the main effect of Level 1 Amount of PS on performance remained significant (Easy: *p*s ranged from 0.001 to 0.007, Hard: *p*s ranged from <0.001 to 0.002).

In sum, the results of our *Correlational* analyses replicate the key finding from Guo and Dobkins (2023) that within-person private speech amount positively predicts visual-spatial working memory performance, and that this effect generalizes to a version of the task made more difficult by virtue of using a set of harder-to-label images. As we discussed in our previous study, finding a positive correlation between amount of private speech and performance still leaves open the question of causality and the direction of causality. Just like the argument we made in our previous study, our analysis of the *content* of participants' private speech is suggestive of greater amounts of private speech leading to greater performance, rather than vice

versa. Still, the most direct way to establish causality is to employ an *experimental* approach, comparing performance between the Quiet and Private Speech conditions, which we address in our *Manipulation* analyses, below.

Manipulation Analyses: Testing the Causal Relationship between Private Speech and Performance

As a starting point, we looked at the effects of both Speech condition (Quiet vs. Private Speech) and Labelability condition (Easy vs. Hard) within the same model, *without* the inclusion of amount of PS. Using a Type III sum of squares multilevel regression model, the dependent variable was performance and the predictor terms were: 1) Speech condition (Private Speech vs. Quiet, entered as a fixed effect), 2) Labelability condition (Easy vs. Hard, entered as a fixed effect), and 3) the interaction between (1) and (2), with Participant included as a random intercept effect. The results of this model revealed the following. *First*, and most germane to our test of causality, participants performed significantly better in the Private Speech, as compared to the Quiet, condition (β = 0.21, 95% CI = [0.08 - 0.35], p = 0.003), which indicates that the use of private speech enhances performance. *Second*, serving as a confirmation that our two sets of images (easy- vs. hard-to-label) create different levels of task difficulty, participants performed significantly better in the Easy, as compared to the Hard, condition (β = 0.59, 95% CI = [0.45 - 0.73], p < 0.001). *Third*, mirroring what we found in the *Correlational* analyses (above), there was no significant interaction between the Speech and Labelability conditions (p = 0.387), indicating that the benefit of private speech manipulation on performance did not differ across the two levels of difficulty. Because the interaction term was non-significant, we removed it and re-ran the model, with the results presented in Table 3.4. This slightly increased the effect size of

the Speech condition (β = 0.26, 95% CI = [0.16 – 0.36], $p < 0.001$) and the Labelability condition (β = 0.63, 95% CI = [0.54 – 0.73], $p < 0.001$). For illustrative purposes, Figure 3.3 visualizes the model-estimated[30] mean performances as a function of Speech and Labelability condition.

[30] A multilevel model considers the nested nature of observations (e.g., trials within individuals) and the variances attributed to between-person differences, inferred from the Intraclass Correlation Coefficient (ICC). By doing so, it estimates the effects while accounting for shared characteristics within each level, facilitating a more accurate examination of the main effects.

Table 3.4: The Results of A Type III Multilevel Model for Testing the Effects of Speech Manipulation and Labelability Manipulation on Performance

	Performance		
Predictors	*β*	*95% CI*	*p*
(intercept)	0.42	0.28 – 0.57	< **0.001**
Speech	0.26	0.16 – 0.36	< **0.001**
Labelability	0.63	0.54 – 0.73	< **0.001**
Random Effects			
$\sigma 2$	0.53		
$\tau 00_{\text{Participant}}$	0.39		
ICC	0.43		
$N_{\text{Participant}}$	106		
Observations	840		
Marginal R^2 / Conditional R^2	0.113 / 0.492		

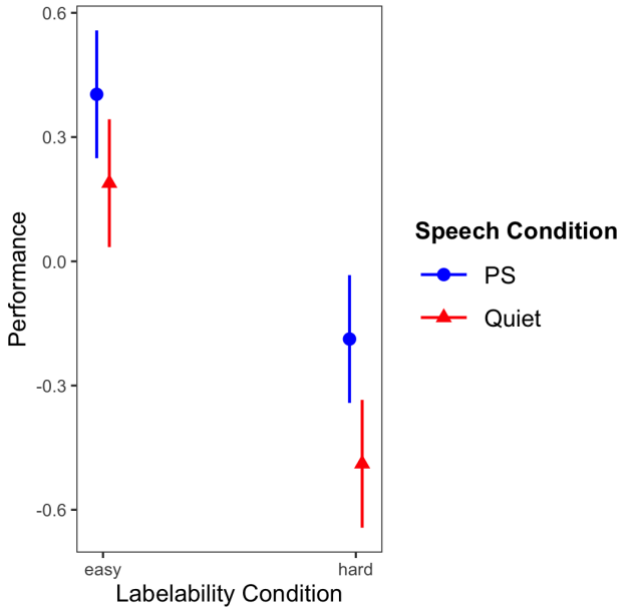

Figure 3.3: The Model-Estimated Mean Performance as A Function of Speech and Labelability from a Type III Multilevel Model.

Note. Performance values are z-scored (see *Methods*), and error bars represent 95% confidence intervals.

Does Level 2 Amount of PS Moderate the Benefit of Private Speech on Performance?

Using Type III sum of squares multilevel regression model, we asked whether between-person (i.e., Level 2) Amount of PS moderates the degree of benefit of private speech on performance observed in the previous analysis. That is, we asked whether the benefit of private speech was greater in participants who used more private speech, which, given the results of the previous analysis, makes intuitive sense would be the case. This analysis was conducted separately for the Easy and Hard conditions. As a first step, Level 2 Amount of PS was calculated by averaging the utterances/min between the two PS trials for each participant (followed by Z-scoring), separately for the Easy vs. Hard conditions. The dependent variable was performance and the predictor terms were: 1) Speech condition (Private Speech vs. Quiet, entered as a fixed effect), 2) Level 2 Amount of PS (entered as a fixed effect), and 3) the

interaction between (1) and (2), entered as a fixed effect, with Participant included as a random intercept effect.

The results of our analyses are shown in Table 3.5 (*Left* panel: Easy, *Right* panel: Hard). Since the direction and effect size of the predictors' coefficients were largely consistent between the Easy and Hard conditions, we present a single narrative for both as follows. The results revealed a significant interaction between Level 2 Amount of PS and Speech condition (Easy: β = 0.25, 95% CI = [0.11, 0.39], p = 0.001; Hard: β = 0.21, 95% CI = [0.06, 0.36], p = 0.005), which, unsurprisingly, was driven by participants who talked out loud more (i.e., higher Level 2 Amount of PS) showing the biggest benefits. To further investigate what pair-wise effects drove these interactions, we conducted post-hoc comparisons, and a visual depiction of the resulting model-estimated performance means is presented in Figure 3.4 (*Panel A*: Easy, *Panel B*: Hard). Post-hoc analyses revealed that Level 2 Amount of PS positively predicted performance in the Private Speech condition (Easy, β = 0.25, 95% CI = [0.10, 0.39], p = 0.001; Hard, β = 0.33, 95% CI = [0.20, 0.47], $p < 0.001$), but not in the Quiet condition (Easy, p = 0.261; Hard, p = 0.623). That is, participants who talked out loud the most outperformed those who talked less, but only on the Private Speech, not the Quiet, condition. The result seems intuitive and also shows that baseline competency on the task (represented by the Quiet trials) does not differ between those who talk more vs. less. We return to a deeper interpretation of this finding in the *Discussion.*

Table 3.5: The Results of A Type III Multilevel Model for Testing the Effects of Level 2 Amount of PS on the Influence of Speech Manipulation on Performance, in the Easy (left) and Hard (right) conditions

	Performance (standardized number of turns)					
	Easy			Hard		
Predictors	*β*	*std.95% CI*	*p*	*β*	*std.95% CI*	*p*
(intercept)	0.19	0.05 – 0.33	**0.008**	0.19	0.04 – 0.33	**0.015**
Speech condition [Private Speech]	0.24	0.10 – 0.38	**0.001**	0.27	0.13 – 0.42	**<0.001**
Level 2 Amount of PS	0.33	0.19 – 0.47	**<0.001**	0.25	0.10 – 0.40	**0.001**
Speech condition [Private Speech] × Level 2 Amount of PS	0.25[31]	0.11 – 0.39	**0.001**	0.21	0.06 – 0.36	**0.005**
Random Effects						
σ2	0.51			0.56		
τ00 Participant	0.26			0.29		
ICC	0.34			0.35		
N Participant	101			101		
Observations	400			401		
Marginal R^2 / Conditional R^2	0.087 / 0.399			0.056 / 0.383		

[31] A positive regression coefficient for the interaction between two variables means the increase of one of the variables increases the effect of the other variable on performance. In this case, as Level 2 Amount of PS increases, the benefit (since the effect is positive) of Private Speech manipulation also increases.

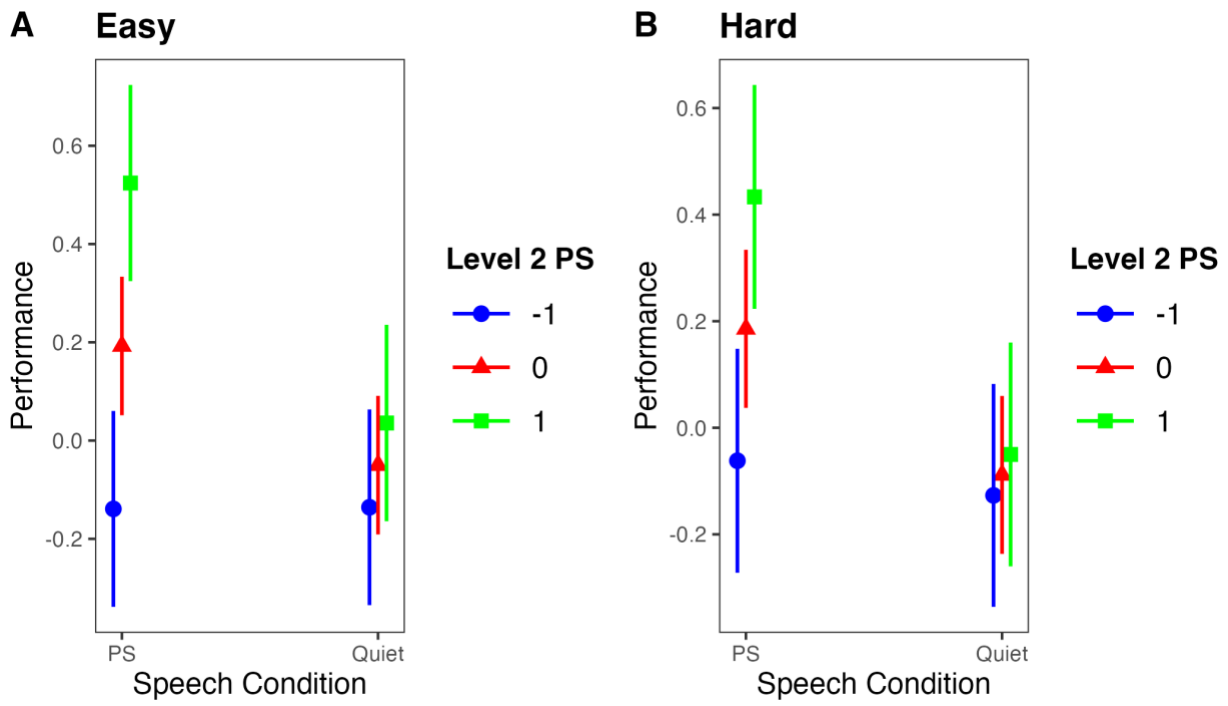

Figure 3.4: The Model-Estimated Performance as A Function of Speech and Level 2 Amount of PS from a Type III Multilevel Model, separately for the Easy (Panel A) and Hard (Panel B) condition

Note. The values "-1", "0", and "1" represent performance levels that are one standard deviation below, at the mean of, or one standard deviation above the Level 2 Amount of PS. The error bars represent 95% confidence intervals.

Does Trait-PS Moderate the Benefit of Private Speech on Performance over and beyond the effect of Level 2 Amount of PS?

In addition to the above analysis, which asked whether participants with higher amounts of private speech show more benefit, it was our intention to ask a similar question regarding trait-PS. Mirroring the spirit of the question asked within the *Correlational* analyses (above), we wondered whether the benefit of private speech manipulation might be greater for people who report more usage of private speech in their everyday life. Like the above analysis, this analysis was conducted separately for the Easy and Hard conditions.

Before proceeding with this question, we wanted to determine whether trait-PS shares variance with Level 2 Amount of PS, since the two constructs could be redundant (i.e., people who report more usage of private speech in their everyday lives are likely to be the same people who produce greater amounts of private speech when instructed to "talk out loud as much as possible" in a laboratory study). To this end, we conducted correlational analyses between Level 2 Amount of PS with all five measures of trait-PS. Although the correlations were low (Easy: *r* values ranging from -0.01 – 0.20, *p*s = 0.068 – 0.936, Hard: *r* values ranging from – 0.06 – 0.21, *p*s = 0.036[32] – 0.805), we deemed it safer to keep Level 2 Amount of PS (and its interaction with Speech condition) in our models testing the moderating effects of Trait-PS, so that the results would reveal effects "over and beyond" those explained by effects of Level 2 Amount of PS, seen above. As such, we conducted five multilevel regression models, separately for each of the five Trait-PS measures, with the dependent variable being performance and the predictor terms being: 1) Speech condition (Private Speech vs. Quiet, entered as a fixed effect), 2) Level 2 Amount of PS (entered as a fixed effect), 3) the interaction between (1) and (2), 4) Trait-PS (entered as a fixed effect), 5) the interaction between (1) and (4), with Participant included as a random intercept effect.

The results of our analyses are shown in Table 3.6 (*Panel A*: Easy, *Panel B*: Hard). Since the direction and effect size of the predictors' coefficients were somewhat consistent between the Easy and Hard conditions, we present a single narrative for both as follows. The results show that for one of the trait-PS metrics, specifically, *Self-Management*, there was a moderating effect on the benefit of private speech on performance. This effect was significant for the Easy

[32] The only significant correlation was between Self-Management subscale of STS and Level 2 Amount of PS, $r(99) = 0.21$, $p = 0.036$.

condition (β = 0.18, 95% CI = [0.04, 0.32], p = 0.010)[33], which was driven by participants who reported more self-management private speech in their everyday lives showing the biggest benefits. To further investigate what pair-wise effects drove this interaction, we conducted post-hoc comparisons, and a visual depiction of the resulting model-estimated performance means is presented in Figure 3.5. Post-hoc analyses revealed that Self-Management negatively predicted performance in the Quiet condition (Easy: β = -0.25, 95% CI = [-0.39, -0.10], p = 0.001, Hard: β = -0.18, 95% CI = [-0.34, -0.03], p = 0.021), but not in the Private Speech condition (Easy: p = 0.357, Hard: p = 0.314). That is, participants who reported using more Self-Management self-talk *under*performed those who reported less, but only in the Quiet, but not the Private Speech, condition. At first glance, this result seems a bit counterintuitive, however, we believe it reflects a "suppression effect". That is, participants who are in the habit of talking out loud to themselves in everyday life may have felt hindered in the Quiet condition where they were explicitly told to keep quiet, an issue we return to in the *Discussion.*

[33] Although the moderating effect of Self-Management in this analysis for Hard was not significant, the directions of the predictors in Hard were the same (compare the Left and the Right panel of Table 3.6).

Table 3.6: The Results of A Type III Multilevel Model for Testing the Effects of Self-Management (Trait-PS) on the Influence of Speech Manipulation on Performance in the Easy (left) and Hard (right) condition

	Performance (standardized number of turns)					
	Easy			Hard		
Predictors	*β*	*std.95% CI*	*p*	*β*	*std.95% CI*	*p*
(intercept)	0.19	0.05 – 0.33	**0.007**	0.18	0.04 – 0.33	**0.014**
Speech condition [Private Speech]	0.25	0.11 – 0.38	**0.001**	0.28	0.13 – 0.42	**<0.001**
Level 2 Amount of PS	0.34	0.20 – 0.48	**<0.001**	0.26	0.11 – 0.42	**0.001**
Self-Manage	-0.06	-0.08 – 0.20	0.368	-0.08	-0.07 – 0.23	0.300
Speech Condition [Private Speech] × Level 2 Amount of PS	0.22	0.08 – 038	**0.002**	0.19	0.04 – 0.34	**0.015**
Speech Condition [Private Speech] × Self-Mange	0.18	0.04 – 0.32	**0.010**	0.10	0.05 – 0.25	0.176
Random Effects						
$\sigma 2$	0.51			0.56		
$\tau 00_{\text{Participant}}$	0.26			0.29		
ICC	0.34			0.35		
$N_{\text{Participant}}$	101			101		
Observations	400			401		
Marginal R^2 / Conditional R^2	0.087 / 0.399			0.056 / 0.383		

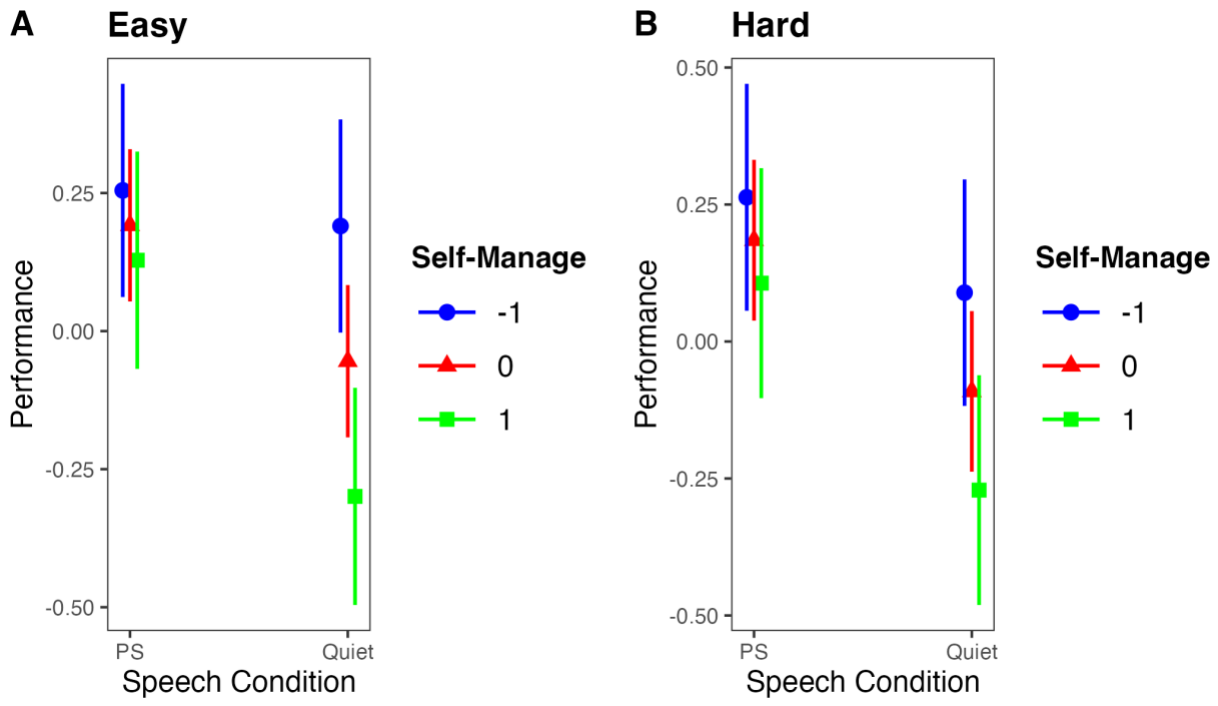

Figure 3.5: The Model-Estimated Performance as A Function of Speech and Trait-PS from a Type III Multilevel Model, separately for the Easy (Panel A) and Hard (Panel B) condition

Note. The values "-1", "0", and "1" represent performance levels that are one standard deviation below, at the mean of, or one standard deviation above the Self-Management subscale of the Self-Talk Scale. The error bars represent 95% confidence intervals.

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 our *Correlational* analyses, within-person performance was compared between two Private Speech trials, in which they were instructed to finish the game in as few turns as possible, while talking out loud to themselves as much as possible. The results of these analyses showed that individuals perform significantly better on trials for which they produce a greater amount of private speech, providing a direct replication of our previous study (Guo and Dobkins, 2023). These correlational findings were boosted by our *Manipulation* analyses, which compared the within-person performance between

the Private Speech condition and a Quiet condition, in which participants were explicitly instructed to *not* talk out loud. The results of these analyses provided direct evidence that the use of private speech *improves* performance. In addition, we found that the effects of private speech (in both our *Correlational* and *Manipulation*, analyses) were comparable between conditions that made the task Easy (with easy-to-label images) vs. Hard (with hard-to-label images), suggesting that the benefits of private speech are invariant across different task difficulties. Finally, we explored the *content* of private speech and found that labeling the images was the most common type of private speech, although, unsurprisingly, this proportion was somewhat lower in the Hard vs. Easy condition.

Another aspect of the current study investigated whether the main findings were *moderated* by between-person (Level 2) variables. Of particular interest was whether Level 2 *amount of PS* moderated the effect of Speech condition in our *Manipulation* analyses[34]. Here, we found that the benefit of private speech was stronger for people who produced greater amounts of private speech. This effect should not be attributable to a "third variable" explanation (for example, if it were the case that more intelligent people both talk out loud more and perform better on the task), since the Manipulation analysis measures within-person *improvement* between the Private Speech and Quiet trials. Post-hoc analyses revealed that this effect was driven by people who talked more outperforming those who talked less, in the Private Speech condition, with no differences in the Quiet condition (see Figure 3.4). At first glance, it seems obvious that the explanation for this result is simple; given the beneficial effects of private speech on performance, people who talk the most when instructed to do so will receive the most

[34] Although we could have tested Level 2 amount of Private Speech in our *Correlational* Analyses, we felt it more direct to test within the *Manipulation* analysis, as this was the litmus test for showing causal effects of private speech on performance.

benefit. This interpretation aligns with what was observed in our *Correlational* analyses, which showed that participants performed better on Private Speech trials in which they produced greater (vs. lesser) amounts of private speech.

However, a more nuanced explanation of the moderating effect of Level 2 amount of private speech might have to do with variations across participants in how difficult they find talking out loud. For example, imagine a participant who finds it difficult to talk out loud when instructed to do so. In addition to the difficult nature of talking out loud resulting in low amounts of private speech for this participant, the additional task of talking out loud (on top of performing the card-matching game) might negatively affect their performance (see Jackson et al., 2023, Rhodes et al., 2019 for evidence that dual-tasks impair memory performance). We suggest that one approach to investigate this possibility is to see if there are participants whose performances are reliably hurt by private speech manipulation compared with control.

Another Level 2 variable we investigated for moderating effects was "Trait Private Speech" (Trait-PS), which were self-report measures of the frequency of using private speech in everyday life. Here, we wondered whether the effects of private speech on performance might be greater for people who habitually talk out loud to themselves in their everyday life. In our *Correlational* analyses, we found no moderating effects of Trait-PS, which indicates that the better performance observed on Private Speech trials that contained higher (vs. lower) amounts of private speech did not depend on Trait-PS. However, in our *Manipulation* analyses, we found that better performance observed in the Private Speech, as *compared to* the Quiet, condition, was moderated by Trait-PS. Specifically, the benefit of using private speech was greater for people who reported using higher amounts of "Self-Management" private speech in their everyday lives. Interestingly, post-hoc analyses revealed that this effect was driven by differences in the Quiet,

but not the Private Speech, condition. Specifically, in the Quiet condition, people who reported using more self-management self-talk *under*performed those who reported using less self-talk, with no differences in the Private Speech condition (see Figure 3.5)[35]. This suggests that participants who habitually use private speech for self-management purposes might be relatively *hindered* in the card-matching game when instructed to "keep quiet". This notion is the "flip side" to the possibility (outlined above) that some participants might be relatively hindered by the additional instruction of talking out loud if they find it difficult to do so. We return to the implications of these findings, below.

A final Level 2 variable we explored for moderating effects was baseline competency, which was computed as the average performance in the two Quiet trials. Mirroring what we did in Guo & Dobkins (2023), we asked whether the positive relationship observed between amount of private speech and performance in our *Correlational* analyses was stronger for participants with lower baseline competency. Replicating our previous study, the current study found no evidence of a moderating effect. Given that baseline competency can be considered a proxy for how difficult one finds the card-matching game (i.e., low competency reflects greater difficulty), the lack of a moderating effect suggests that the effects of private speech on performance may not vary across different levels of (perceived) task difficulty. At first, this null result might be perplexing as previous studies that look at the effects of *verbalizing out loud* (which shares some phenomenology with private speech, see Guo and Dobkins, 2023) have shown that the effects of verbalizing are greater for people who start out at lower competency levels. For example, Kray et

[35] In spirit, comparing performance between two Private speech trials (which vary in their amount of private speech), as in our *Correlational* analysis, should be similar to comparing performance between the Private Speech and Quiet trials (which *also* vary in their amount of private speech, with the latter having 0), as in our *Manipulation* analysis. The finding that Trait-PS moderated the effects of private speech *only* in the Manipulation analysis is due to the fact that the moderation effect was driven by variations across people in the Quiet, and not the Private Speech, condition.

al. (2008) investigated the benefits of verbalization on cognitive performance across different developmental stages (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). Most relevant to the current discussion, they found that the benefit of task-relevant speech was greatest for the two age groups (young children and older adults) with the *lowest* baseline performance. In the current study, participants were in the *same* age group (i.e., young adults). As such, rather than interpreting our null result as evidence that the effects of private speech on visual-spatial working memory performance do not depend on task difficulty, it may be that we did not have sufficient variation in perceived task difficulty across our cohort of participants to show an effect.

Rather than hoping a single version of the card-matching game varies sufficiently in difficulty across participants to see the effect of difficulty (as in the developmental literature, above), a more direct approach is for the experimenter to create different versions of the task, which – with a single participant – vary in difficulty. Although correlational in nature, one study that took this approach in children (five- to six-year-olds) was Fernyhough and Fradley (2005). They manipulated the number of moves necessary to finish an executive function task (with

difficulty Levels 1 to 4) and discovered an inverted U-shaped relationship between task difficulty and the amount of (spontaneous) private speech produced, meaning that children's private speech frequency peaked at medium difficulty (Levels 2 and 3). In the current, we likewise manipulated task difficulty through the creation of easy- vs. hard-to-label images. Corroborating previous literature showing the effects of labelability on memory performance (see *Introduction*) our participants performed significantly worse when the images were Hard- vs. Easy-to-label, and from this, we surmise that the former condition was more difficult for participants. Because we found that the effects of private speech were comparable between the Easy vs. Hard condition, we are left with the same two possibilities discussed above when asking if baseline competency moderates the effects of private speech: either the effects of private speech on visual-spatial working memory performance do not depend on task difficulty, or our labelability manipulation did not produce enough variation in task difficulty to witness its effects. Future studies might try to create more pronounced differences in difficulty, e.g., by scrambling images of real-life objects to make them completely meaningless and hard to label.

Implications and Future Directions

The findings of the current study show that talking out loud — when instructed to do so, improves cognitive performance in adults. Given these observed benefits, and their implications for real-world educational/instructional settings, future studies should consider other variables that might moderate/enhance the private speech benefit. The current study shows that Level 2 amount of private speech is an important moderating variable; as might be expected, people who talk out loud more receive greater benefits. We also found that *restraining* oneself from using private speech can actually impair cognitive performance for people who habitually use it for

self-management in everyday life. To the extent that we could test it, our results also show that the benefits of using private speech generalize across task difficulty.

An interesting future direction will be to investigate the effects of different types of private speech on performance, by explicitly instructing participants to use different types (e.g., in the verbalization literature, see Souza et al., 2021). Although visual inspection of the content of participants' private speech in the current study showed some differences between the easy-to-label vs. hard-to-label conditions (Table 3.2), a study like ours that only describes the content of participants' private speech cannot draw conclusions regarding the causal relationship between content and performance; that is, people might spontaneously (and intuitively) use the type of private speech that is optimal for a given task, or a given task might bias people to use a certain kind of private speech (e.g., see Mulvihill et al., 2021). In fact, one explanation for why we saw equal benefits of private speech in our Easy vs. Hard condition is that our participants intuitively knew to adapt the content of their speech for optimal performance. Moreover, it may not be that in all situations, the more private speech, the better. For example, in the current study, the improved performance in the Private Speech vs. Quiet condition was comparable between the Easy and Hard condition, despite the fact that the rate of utterances was significantly higher for the former.

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). 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 (across different developmental stages) 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.

Acknowledgments

Chapter 3, in full, has been submitted for publication of the material as it may appear in Consciousness and Cognition 2023. Xinqi, Guo; Dobkins, Karen, Elsevier. The dissertation author was the primary researcher and author of this paper.

Appendix C

Trial order and block order

Although not part of our pre-registration, we thought it wise to examine the effects of "Trial Order" and "Block Order". As explained in the *Methods*, the order of speech manipulation was counterbalanced across participants, with half of the participants starting with the Quiet trials first (Trial Order = 0), and the other half of the participants starting with the Private Speech trials first (Trial Order = 1), and for each participant, Trial Order was maintained across the two labelability conditions. The order of labelability manipulation was counterbalanced across participants, with half of the participants starting with the Easy first (Block Order = 0), and the other half of the participants starting with the Hard first (Block Order = 1), and for each participant.

The analyses here focused on exploring the potential effects of Trial Order and Block Order on results reported in the main text of the study. By examining these order effects, the analyses provided in Appendix C aim to understand the potential complexity of doing experiments on private speech, and how the arrangement of conditions may influence the insights can be drawn from the experiments.

Correlational Analyses

Does Trial Order Moderate the Effect of Level 1 Amount of PS on Performance

Specifically, we asked whether Trial Order moderates the positive relationship between Level 1 PS and performance observed in the prior analysis. Using a Type III sum of squares multilevel regression model, separately for the Easy vs. Hard condition, the dependent variable was performance and the predictor terms were: 1) Level 1 Amount of PS (entered as a fixed effect), 2) Trial Order (entered as a fixed effect), and 3) the interaction between (1) and (2), with Participant included as a random intercept effect. Although we found no significant interactions,

there was a main effect of Trial Order in the Easy condition (β = 0.36, 95% CI = [0.02, 0.70], p = 0.040). This was driven by the fact that participants who were tested with the Quiet trials first performed better than participants who were tested with the Private Speech trials first. We speculate that this may be due to the fact that when tested with the Quiet trials first, participants got a chance to become familiarized with the card-matching task before they were asked, in the subsequent trials, to perform an *additional* task (i.e., talk out loud as much as possible) on top of the card-matching task. This finding should be interpreted cautiously, however, for two reasons. First, it was observed only in the Easy, and not the Hard, condition, whereas our intuition would predict, if anything, the opposite. Second, Trial Order is a complicated construct in the current design (see *Methods*).

Does Block Order Moderate the Effect of Level 1 Amount of PS on Performance

Specifically, we asked whether Block Order moderates the positive relationship between Level 1 PS and performance observed in the correlational analysis. Using a Type III sum of squares multilevel regression model, separately for the Easy vs. Hard condition, the dependent variable was performance and the predictor terms were: 1) Level 1 Amount of PS (entered as a fixed effect), 2) Block Order (entered as a fixed effect), and 3) the interaction between (1) and (2), with Participant included as a random intercept effect. Although we found no significant interactions, there was a main effect of Block Order regarding performance in the Hard condition (β = -0.46, 95% CI = [-0.79, -0.14], p = 0.006). This was driven by the fact that participants who were tested with the Easy images first performed better in the subsequent Hard condition than the participants who were tested with the Hard images first. We speculate that this may be due to the fact that when tested with the Easy images first, participants had a chance to become

familiarized with the card-matching task before they were asked, in the subsequent block, to do the card-matching task on a more difficult set of images.

Manipulation Analyses

Trial Order

Specifically, we asked whether Trial Order moderates the benefit of private speech relative to the quiet condition observed in the manipulation analysis. Using a Type III sum of squares multilevel regression model, separately for the Easy vs. Hard condition, the dependent variable was performance and the predictor terms were: 1)Speech condition (entered as a fixed effect), 2) Trial Order (entered as a fixed effect), and 3) the interaction between (1) and (2), with Participant included as a random intercept effect. There was no significant interaction between Trial Order and Speech condition, indicating that the private speech benefit was not moderated by Trial Order. Note that the benefit of private speech manipulation in the Hard condition was only marginally significant after adding Block Order, possibly due to having more predictors in the same model.

Block Order

Specifically, we asked whether Block Order moderates the benefit of private speech manipulation observed in the manipulation analysis. Using a Type III sum of squares multilevel regression model, separately for the Easy vs. Hard condition, the dependent variable was performance and the predictor terms were: 1) Speech condition (entered as a fixed effect), 2) Block Order (entered as a fixed effect), and 3) the interaction between (1) and (2), with Participant included as a random intercept effect. Although we found no significant interactions, there was a main effect of Block Order regarding performance in the Hard condition (β = -0.34,

95% CI = [-0.66, -0.01], $p = 0.044$). We believe the interpretation of the main effect of Block Order is the same as the interpretation of the effect of Block Order for the correlational analysis and will not repeat the interpretation again here.

GENERAL DISCUSSION

Humans might be the only animal species that is capable of self-talk. Adults have two forms of self-talk: inner speech (inaudible) and private speech (audible). The dominant theory of self-talk regards private speech as a typical and adaptive behavior during middle childhood when inner speech is not yet mature. As individuals transition past middle childhood, private speech becomes less frequent, with inner speech taking prominence.

This dissertation fills a significant gap in the literature, providing evidence of the important and adaptive role of private speech in adult cognition. Additionally, when prompted, their cognitive performance improves with the use of private speech during tasks. Intriguingly, a specific subgroup of young adults – those who habitually employ private speech in daily cognitive tasks – performed worse when asked to remain silent.

I believe the dissertation lays the groundwork for highlighting the advantages of private speech in adult cognition, which warrants further replication and research. Should the advantages of private speech be consistently observed, it would be beneficial to encourage adults to incorporate private speech into their daily cognitive routines.

A potential critique of this study pertains to the limited exploration of internal processes in conditions without private speech. While one of my initial aims was to contrast the effects of private speech and inner speech on performance, Chapter 1's findings cast doubt on the reliability of self-reported self-talk – the most practical method to measure concurrent inner speech. Consequently, Chapters 2 and 3 focused on private speech. While the extent of inner speech use remained uncertain, the observed advantages of the private speech condition can be dissected into the following potential scenarios:

a) If participants utilized inner speech in the control condition as much as, or more than, the private speech in the private speech condition, then the observed benefits suggest private speech is more advantageous to adult cognition than inner speech.
b) If participants used less inner speech in the control than the private speech in the private speech condition, the observed benefits don't conclusively compare the efficacy of private speech to inner speech. This is because the benefit seen in the private speech condition could be due to private speech being more, equally, or even less effective than inner speech, yet still resulting in a comparative advantage between conditions.

However, these scenarios become important when the goal is to directly compare the efficacy of private speech and inner speech. Such a direct comparison, beyond the scope of this dissertation, might necessitate a more precise quantification of inner speech.

Future studies

Insights from this dissertation generate ample opportunities to further the research of private speech on cognitive performance and have implications in both theory advancement and guidance for practitioners like educators as well as students' self-guidance. Below are studies ideas that could be low-hanging fruits as the next rounds of investigation.

In Chapter 3, I started to explore the potential individual differences that might make private speech especially helpful. We found that the benefit of private speech manipulation was the most pronounced for individuals who habitually use private speech in their everyday lives. A deeper understanding of the effects of private speech on adult cognition can benefit from incorporating a broader range of age groups.

During normal aging, cognitive capacities decay at different rates (Murman, 2015). The most evident cognitive declines in older individuals are observed in tasks that require fast information processing and decision-making, such as speed of processing, working memory, and

executive function. On the other hand, speech and language functions largely remain intact, even into advanced age (Bigler, 2012). Given this, how older adults typically leverage private speech or how they might benefit from it as a verbal strategy to compensate for other relatively diminished capacities deserves further exploration.

One specific avenue to investigate is the potential of private speech to reduce common errors made by the elderly during cognitive tasks, like perseveration errors. Perseveration errors refer to the repetition of action patterns that are no longer adaptive or beneficial. For instance, in the card-matching game used in the dissertation, a perseveration error would occur if a participant persistently reveals a card with a pattern they have already seen and identified as a non-match. To illustrate, if a participant uncovers Card A and sees a “star” pattern and then later exposes Card B to find a “moon” pattern, a perseveration error would be made if they then reveal Card A again, recalling its pattern but without any strategic intention (Eppinger et al., 2011; Head et al., 2009). By prompting private speech during such tasks and observing whether it diminishes errors like perseveration, we can evaluate whether private speech serves an adaptive function in aging.

Finally, there should be a continuation of research into the mechanism of how private speech aids cognition.

One possibility is that participants might not have been sufficiently attentive when they weren't prompted to use private speech. In simpler terms, rather than making items more distinctive in memory, private speech might merely enhance task attention. This implies that the non-private speech condition might made participants less attentive to the game. To verify whether the benefits observed in the private speech group can be attributed to a lack of attention in the non-private speech group, a couple of approaches can be adopted:

Firstly, a control condition can be established where participants are directed to say "this" every time they reveal a card. Words like "this" or “that” have been frequently used between children and caregivers to reference objects and to guide attention. This usage persists into adulthood, making it a potentially effective tool to match attention levels across participants. However, it's worth noting that some participants already used words like “this” or “that” in their private speech. These words were often used either just before or during the action, with the former hinting at self-guidance and the latter being more of a concurrent behavior. Another approach to address this attention account is to provide small incentives to the silent group, ensuring they pay equal attention to the task.

To determine if private speech has an impact on memory specifically, a recognition test can be incorporated post-game. Here, participants would differentiate between "Old" (cards they've seen during the game) and "New" (cards they haven't encountered) patterns. This stimuli pool for the recognition test would include patterns from the game as well as new, perceptually similar patterns. By analyzing both correct and false identification rates, we can ascertain if the overall memory efficiency (reflected in fewer turns to match hidden pairs) in the private speech condition stems from better memory for the patterns or a more conservative approach to revealing cards during the task.

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.

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.

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.

Allen, P., Aleman, A., & Mcguire, P. K. (2007). Inner speech models of auditory verbal hallucinations: Evidence from behavioural and neuroimaging studies. International Review of Psychiatry, 19(4), 407–415.

Asp, I. E., Störmer, V. S., & Brady, T. F. (2021). Greater visual working memory capacity for visually matched stimuli when they are perceived as meaningful. Journal of cognitive neuroscience, 33(5), 902-918.

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. Annual Review of Psychology, 63(1), 1–29.

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.

Behrend, D. A., Rosengren, K., & Perlmutter, M. (1989). A new look at children's private speech: The effects of age, task difficulty, and parent presence. International Journal of Behavioral Development, 12(3), 305-320.

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.

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.

Bigler, E. D. (2012). Symptom Validity Testing, Effort, and Neuropsychological Assessment. Journal of the International Neuropsychological Society, 18(4), 632–640.

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. Proceedings of the National Academy of Sciences, 113(27), 7459–7464.

Brady, T., & Störmer, V. (2021). The Role of Meaning in Visual Working Memory: Real-World Objects, But Not Simple Features, Benefit From Deeper Processing. Journal of Experimental Psychology: Learning, Memory, and Cognition.

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.

Diaz, R. (2014). Methodological Concerns in the Study of Private Speech. In Private Speech (pp. 65–92). Psychology Press.

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.

Dunbar, K., & Sussman, D. (1995). Toward a Cognitive Account of Frontal Lobe Function: Simulating Frontal Lobe Deficits in Normal Subjects. Annals of the New York Academy of Sciences, 769(1 Structure and), 289–304.

Duncan, R. M., & Cheyne, J. A. (1999). Incidence and functions of self-reported private speech in young adults: A self-verbalization questionnaire. Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement, 31(2), 133–136.

Duncan, R. M., & Cheyne, J. A. (2001). Private speech in young adults: Task difficulty, self-regulation, and psychological predication. Cognitive Development, 16, 889–906.

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.

Eppinger, B., Hämmerer, D., & Li, S.-C. (2011). Neuromodulation of reward-based learning and decision making in human aging. Annals of the New York Academy of Sciences, 1235(1), 1–17.

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.

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.

Fernyhough, C. (2004). Alien voices and inner dialogue: Towards a developmental account of auditory verbal hallucinations. New Ideas in Psychology, 22(1), 49–68.

Fernyhough, C., & Fradley, E. (2005). Private speech on an executive task: Relations with task difficulty and task performance. Cognitive Development, 20(1), 103–120.

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.

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.

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.

Gentner, D., & Goldin-Meadow, S. (Eds.). (2003). Language in mind: Advances in the study of language and thought.

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.

Guerrero, M. C. M. de. (2018). Going covert: Inner and private speech in language learning. Language Teaching, 51(1), 1–35.

Hatzigeorgiadis, A., & Galanis, E. (2017). Self-talk effectiveness and attention. Current Opinion in Psychology, 16, 138–142.

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.

Head, D., Kennedy, K. M., Rodrigue, K. M., & Raz, N. (2009). Age differences in perseveration: Cognitive and neuroanatomical mediators of performance on the Wisconsin Card Sorting Test. Neuropsychologia, 47(4), 1200–1203.

Hurlburt, R. T., Alderson-Day, B., Kühn, S., & Fernyhough, C. (2016). Exploring the Ecological Validity of Thinking on Demand: Neural Correlates of Elicited vs. Spontaneously Occurring Inner Speech. PLOS ONE, 11(2), e0147932.

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.

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.

Langdon, R., Jones, S. R., Connaughton, E., & Fernyhough, C. (2009). The phenomenology of inner speech: Comparison of schizophrenia patients with auditory verbal hallucinations and healthy controls. Psychological Medicine, 39(4), 655–663.

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).

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.

Lupyan, G. (2008). From chair to" chair": a representational shift account of object labeling effects on memory. Journal of Experimental Psychology: General, 137(2), 348.

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. Frontiers in psychology, 3, 54.

MacLeod, C. M., & Bodner, G. E. (2017). The Production Effect in Memory. Current Directions in Psychological Science, 26(4), 390–395.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. Journal of Experimental Psychology: Learning,
Memory, and Cognition, 36(3), 671–685.

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.

Martin, C. F., & Shumaker, R. W. (2022). Orangutan strategies for solving a visuospatial memory task. American Journal of Primatology, 84(10), e23367.

McCarthy-Jones, S., & Fernyhough, C. (2011). The varieties of inner speech: Links between quality of inner speech and psychopathological variables in a sample of young adults. Consciousness and Cognition, 20(4), 1586–1593.

Morin, A. (2012). Inner Speech. In Encyclopedia of human behavior; editor-in-chief, V.S. Ramachandran. (2nd ed.).

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

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.

Murman, D. L. (2015). The Impact of Age on Cognition. Seminars in Hearing, 36(3), 111–121.

Nakabayashi, K., & Mike Burton, A. (2008). The role of verbal processing at different stages of recognition memory for faces. European Journal of Cognitive Psychology, 20(3), 478–496.

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.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. Journal of Neuroscience Methods, 162(1), 8–13.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision [Computer software]. OpenAI.

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.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2013). How does using object names influence visual recognition memory? Journal of Memory and Language, 68(1), 10–25.

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.

Rowe, M. L., & Goldin-Meadow, S. (2009). Early gesture selectively predicts later language learning. Developmental Science, 12(1), 182–187.

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].

Schubert, T. (2022). Labels aid visual working memory. Nature Reviews Psychology, 1(12), Article 12.

Schumann-Hengsteler, R. (1996). Children's and Adults' Visuospatial Memory: The Game Concentration. The Journal of Genetic Psychology, 157(1), 77–92.

Shi, X., Brinthaupt, T., & McCree, M. (2017). Understanding the Influence of Self-Critical, Self-Managing, and Social-Assessing Self-Talk on Performance Outcomes in a Public Speaking Context. Imagination, Cognition and Personality, 36(4), 356–378.

Sobrinho, N. D., & Souza, A. S. (2023). The interplay of long-term memory and working memory: When does object-color prior knowledge affect color visual working memory? Journal of Experimental Psychology: Human Perception and Performance, 49(2), 236–262.

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.

Souza, A. S., & Skóra, Z. (2017). The interplay of language and visual perception in working memory. Cognition, 166, 277–297.

Souza, A. S., Overkott, C., & Matyja, M. (2021). Categorical distinctiveness constrains the labeling benefit in visual working memory. Journal of Memory and Language, 119, 104242.

Tangram Channel—ESSENTIAL CARDS. (n.d.). Providing Teachers and Pupils with Tangram Activities. Retrieved May 20, 2023, from http://www.tangram-channel.com/tangram-cards/essential-cards/

Tullett, A. M., & Inzlicht, M. (2010). The voice of self-control: Blocking the inner voice increases impulsive responding. Acta Psychologica, 135(2), 252–256.

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.

Uttl, B., Morin, A., & Hamper, B. (2011). Are Inner Speech Self-Report Questionnaires Reliable and Valid? Procedia - Social and Behavioral Sciences, 30, 1719–1723.

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).

Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. Journal of Experimental Psychology: General, 147(1), 113–124.

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.

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.

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.