TRANSFER SUCCESS ON THE LINDA PROBLEM: A RE-EXAMINATION USING DUAL PROCESS THEORY, LEARNING MATERIAL CHARACTERISTICS, AND INDIVIDUAL DIFFERENCES

MICHAEL SHIUMING TRUONG

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

August 2022

Abstract

The Linda problem is an intensely studied task in the literature for judgments where participants judge the probability of various options and frequently make biased judgements known as conjunction errors. Here, I conceptually replicated and extended the finding by Agnoli and Krantz (1989) that when participants are explicitly trained with Venn diagrams to inhibit their heuristics, successful transfer of learning is observed. I tested whether transfer success was maintained: (1) when the purpose of the training was obscured; (2) after controlling for individual differences; and (3) when learning materials did not include visual images. I successfully replicated their finding, identifying transfer success when the purpose of the training was masked and after controlling for individual differences. Furthermore, the effects of individual differences on transfer success depends on both the kind of learning material used and whether the purpose was masked. Hence, these findings support claims that education can inhibit biases.

Keywords: Memory; Learning; Transfer; Linda Problem; Representativeness Heuristic; Conjunction Error; Individual Differences

Acknowledgments

Some say that it takes a village to raise a child. This thesis is the cumulation of not only my own efforts, but my experiences with all the people in my life. For all that has happened and that will happen, I owe everything to my village. Even if the oceans were to be emptied and filled with ink, there would still not be enough ink to express my appreciation.

I was first encouraged to study psychology and neuroscience by my highschool science teacher, Rebecca Jeffrey. I still remember the shock on her face when I told her about my encounter with someone who strongly denounced psychology's status as a science. While I was at the University of Toronto for my bachelor's, I benefitted from the teaching and guidance of Drs. Kaori Takehara-Nishiuchi, Romin Tafarodi and Morris Moscovitch. My first experience as a research assistant was in Dr. Takehara-Nishiuchi's lab under the supervision of her PhD student, Maryna Pilkiw, and it was there that I was first struck by the power that a simply-worded and short question could have. I am still in awe at the penetrating depth her questions, all asked with clarity. It was through Dr. Tafarodi's course in critical psychology and many other teaching moments that my style of and focus in critiquing psychological research was first made. I also owe a great intellectual debt to Drs. Paul E. Meehl and André Kukla for their works on the philosophy of psychology, which played an important role in the development of my views of psychological research. Last, but not least, I owe too much to Dr. Moscovitch for his generous guidance, encouragements and example, particularly—but not limited to—while I was supervised by his PhD student, Nichole Bouffard. I would be nowhere near as proficient as I am at R without Nichole's patience. His lab was an extremely intellectually fertile environment and although our first experiment was not successful by the usual standards, it has forever impacted the care and thought that I attempt to put in conceiving and interpreting any experiment.

To my committee and co-supervisors, Drs. Thanujeni Pathman, W. Dale Stevens and Maggie Toplak, thank you for this opportunity and all the encouragement and guidance that led to this work. Like a team of blacksmiths, significant hammering and heating was involved, but also as much or more tender quenching and sharpening. Most importantly, all this was done with great care and the best of intensions. Not everyone can say the same and, most importantly, I am all the better person and scientist for it. In particular, Dr. Pathman was absolutely important in every aspect related to the verbal paired associates task and ensuring that I stay on track. Furthermore, Drs. Pathman and Stevens were very encouraging even when this thesis explored their discomfort zone. Dr. Stevens encouraged and supported many aspects of this thesis that made it a stronger and more interesting work, particularly the use of individual differences to qualify transfer success and the comparison of learning materials with graphs and text versus text only. Lastly, for her expertise, time, great generosity and much more, this thesis owes much to Dr. Toplak—without her this thesis would not exist. Thank you also to my examining committee, Drs. David Flora and Ji Yeh Choi. I hope that the statistical analyses and interpretations done here have made the best use of—but not abuse of—statistics.

To my labmates, Naail, Lily, Mylann, Amir, Taran, Owen, Aarthi, Vicente, Katherine and Braxton, thank you for being there throughout the COVID-19 pandemic and all those weekly chats. Without you, my master's would not have been nearly as interesting as it has been. To all my friends and family, particularly Mom, Dad, Darren, Hugo, Brian, Raymond, Lorenzo, Luke, Ronan, Martin, Ian, Josh and many others, thank you for all those late night conversations where I raved about what I was researching. All of you come from vastly different backgrounds and disciplines, yet you would still listen closely even when I could not be more unclear. The direction and energy behind every expression I make is thanks to you.

The experiment behind this thesis also owes much to many others. Thank you to York University for trusting in my proposal and financing a significant portion of the costs. These costs were also generously financed by Drs. Pathman and Stevens. Thank you to Drs. Ian Clark and Eleanor Maguire for sharing the pool of highly imageable and concrete words that I used for my verbal paired associates task, which originated from their 2018 experiment (I. A. Clark et al., 2018). Thank you to CommonLit for providing the reading comprehension materials—authored by Kubic (2016) and adapted with minimal wording changes—used in this thesis. Thank you to Drs. Franca Agnoli and David H. Krantz for openly sharing the training materials used in this thesis. Their materials are reprinted and adapted here from Appendix A of Cognitive Psychology, 21, Agnoli & Krantz, Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy, 515-550, Copyright Elsevier (1989). I also would not have been drawn to this area of research were it not for Michael Lewis' *The Undoing Project* (Lewis, 2017), a beautifully written story of the friendship and research of Amos Tversky and Daniel Kahneman. This thesis is dedicated to their friendship, to Daniel Kahneman and Amos Tversky.

To the many erudite statistical consultants at York, thank you for your assistance throughout this work. Thank you to Dr. Georges Monette for advising the use of likelihood ratio statistics and binomial regressions for this experiment. Thank you to Dr. Michael Friendly for recommending the in-depth guide on the topic of marginal effects and their visualization by Heiss (2022). Thank you to Dr. Robert Phil Chalmers for many meetings and late night emails for how to allocate participants in this study and specifying the regression models to answer my research questions.

Lastly, although I may feel that this work represents the pinnacle of my reasoning about and studying human psychology, I hope that this is only the beginning of many other insights.

Table of Contents

List of Figures

List of Tables

Transfer Success on the Linda Problem: A Re-Examination Using Dual Process Theory, Learning Material Characteristics, and Individual Differences

Barnett and Ceci (2005) write that "an instructional program that successfully inculcates skills, but for which the skills do not transfer to nonacademic [*sic*] situations outside the classroom, is a failure. … No one cares about learning if it stops at the schoolhouse door" (p. 295). This quote highlights what is at stake in the ongoing controversy regarding the transfer of learning and it is a problem that goes beyond mere academic quibbling. As shown by Kahneman and Tversky's research, even the most educated people can make seriously biased judgments through the misapplication of heuristics (Kahneman, 2011; Tversky & Kahneman, 1974, 1983). Specifically, one of their most disputed (Gigerenzer, 1991; Hertwig et al., 2008; Hertwig & Chase, 1998; Hertwig & Gigerenzer, 1999; Mellers et al., 2001), but also replicated (Agnoli, 1991; Agnoli & Krantz, 1989; Fiedler, 1988; Fisk & Pidgeon, 1997; Hertwig & Chase, 1998; Mellers et al., 2001), findings is how many people make biased judgments on the Linda problem (Tversky & Kahneman, 1983).

In the Linda problem, participants are given a description of a young woman who majored in philosophy, is interested in social justice, and took part in anti-nuclear demonstrations. Then, participants are asked to rank whether it is more likely for her to be an (A) bank teller; (B) feminist; or (A and B) bank teller and feminist (alongside 5 other possibilities). Tversky & Kahneman (1983) found that both statistically-naïve and statistically-educated participants were very likely to rank the conjunction (A and B) as more likely than the unrepresentative constituent (A) (cf. Hertwig & Chase, 1998)[1]—a mistake they called the

[1] Hertwig & Chase (1998) found that when statistical sophistication was measured via background problems related to statistics after completing the Linda problem, then statistical sophistication was positively correlated with

conjunction error (CE). Tversky & Kahneman (1983) reasoned that the probability of someone being A and B cannot be higher than being either A or B because being either A or B also includes the possibility of being A and B. Thus, they argue that choosing the conjunction is a grave error in elementary probability (Tversky & Kahneman, 1983; cf. Gigerenzer, 1991).[2] Through a series of studies, Tversky and Kahneman (1983) attributed the CEs to participants' 'representativeness heuristic': a tendency to rate the most representative options as the most likely. Therefore, they argue that the reason why participants tended to rank the conjunction (bank teller and feminist) as more likely than its unrepresentative constituent (bank teller) is because the conjunction also had a highly representative option (feminist). Although Tversky and Kahneman (1983) accepted that their naïve participants might commit CEs, they were surprised that their educated participants were about as likely to make the same error (cf. Hertwig & Chase, 1998). Whatever statistical training these educated participants had did not transfer to the Linda problem.

This failure in identifying transfer of learning is not unique to the domain of judgments. Currently, there is significant debate regarding whether the experiences learned from 'brain training' games (Dahlin et al., 2008; Nichols et al., 2021; Owen et al., 2010; Simons et al., 2016; von Bastian et al., 2022) or certain kinds of activities (Bialystok, 2017, 2021; Bialystok & Craik, 2022; Meltzer et al., 2021; Nichols et al., 2020; Olszewska et al., 2021) will broadly transfer

'correct' responses on the Linda problem. However, their measure of statistical sophistication may be inappropriate because although statistical training may help in answering their problems, statistical training is not a necessary pre-requisite due to the general nature of their problems. Thus, their measure may be contaminated by measuring numeracy, rather than statistical training itself. Therefore, the finding by Tversky and Kahneman (1983) that statistically-naïve and educated participants perform similarly on the Linda problem may still hold, despite the positive correlation between statistical sophistication, as defined by Hertwig & Chase (1998), and correct responses to the Linda problem.

[2] Gigerenzer (1991) is a landmark critique of Tversky and Kahneman's general line of research. In this paper, Gigerenzer argues that Tversky and Kahneman's "CEs" are not necessarily errors, and that the 'true' normative response is disputable. This being said, I explain in Footnote 3 why the truly correct response to the Linda problem is unimportant for my purposes.

outside of their learned contexts and forms. It is difficult to argue for the importance of learning to cognition, if its benefits are so restricted to their original contexts. However, this dilemma also creates an opportunity. I argue that if it can be shown that formal learning truly can have a significant impact on people's judgments, then this also legitimizes the creation of a theoretical framework where our understanding of memory and judgments are tightly connected. Furthermore, the integration of seemingly disparate domains of research is critical to the progress of scientific theories (Kukla, 2001).

In contrast with the transfer-failure finding of Tversky and Kahneman (1983), Agnoli and Krantz (1989) found that if they trained naïve participants in the rules of probability related to conjunction errors immediately before answering a set of Linda and Linda-like problems, then trained participants made significantly fewer errors than untrained participants in a control group. On average, trained participants made CEs on about 44% of the problems, whereas the control group made CEs on about 73% of the problems (Agnoli & Krantz, 1989)—an improvement which has been conceptually replicated at least twice (Agnoli, 1991; Fisk & Pidgeon, 1997). Thus, these findings have been cited to argue that education may inhibit the effects of heuristics and biases (Agnoli, 1991; Agnoli & Krantz, 1989; Kahneman, 2003), an effect which I will dub 'transfer success' hereafter.[3]

However, the recent proliferation of the dual process theory of judgments in the field of decision-making (Kahneman, 2011) provides an impetus for revisiting claims of transfer success.

[3] Throughout this thesis, I operationally define correct responses to the Linda problem as instances of when the conjunction is ranked less probable than either of its constituents—i.e., when participants do not make CEs. I do not claim that CEs are the normative incorrect response. Instead, I argue that the true correct response is unimportant to the central motivation of my proposed experiment. This experiment focuses on the match and mismatch between what people are taught and how they will behave whether transfer occurs or not. This experiment focuses on how if we assume that CEs are incorrect, then the published claims of transfer success on the Linda problem are incommensurate with dual process theory. Hence, I seek to amend the situation through a more comprehensive experiment.

Specifically, what most prior and even recent research on transfer success to problems of heuristics and biases seem to target is whether people can remember and apply rules (of probability) to their judgments, but I argue that this is not the most significant problem through the perspective of dual process theory. According to dual process theory, the real problem is that under certain circumstances: (1) participants fail to recognize that their learning should be applied to their judgments; and/or (2) participants make a fast and intuitive judgment using system 1, which overrides any further careful thinking using system 2 (Kahneman, 2011). (The terms system 1 versus 2 thinking generally refer to fast and intuitive versus slow and careful thinking, respectively.) What this means is that—in our educational context—to test transfer success, we need to be confident that what we are testing is that participants are more likely to naturally recognize that their learning should be applied to a problem, then apply it correctly. Hence, to my knowledge, there currently is no experimental evidence for transfer success on dual process theory's terms because prior studies either (1) had an experimental context that strongly implied to participants that they should apply what they learned to the Linda problem (e.g. Agnoli, 1991; Fisk & Pidgeon, 1997); or (2) explicitly instructed them to do so (Agnoli & Krantz, 1989). When participants are cued to apply their learning to the Linda problem before they have even read the problem, then there is no need for a natural recognition *that* their learning should be applied to the problem. Participants need only follow the instructions for experimenters to observe a change in performance (Barnett & Ceci, 2002; Jones, 2009b). Furthermore, in real-world contexts, there rarely is an explicit external prompt to apply one's learning to an upcoming problem. Therefore, it is still unknown whether learning can inhibit the heuristics that people typically use to make biased judgments and lead to fewer conjunction

errors on the Linda problem, when the problem appears ‘naturally’, without strong cues to the desired or correct response.

To address this gap in our empirical base, here, I conceptually replicate the first experiment of Agnoli and Krantz (1989) and extend their findings with an additional condition where the relationship between the learning materials and the Linda problem is masked, an implicit condition. Moreover, establishing transfer success on dual process theory’s terms is not the only way my extension of their experiments seeks to improve our understanding of transfer success to problems of heuristics and biases.

More than 30 years have passed since the publication of the original experiment by Agnoli and Krantz (1989), and since then, a long line of research has identified dispositional individual differences that predict how people will respond to judgments involving heuristics and biases (Stanovich, 2000, 2016; Stanovich & Toplak, 2016b; Toplak et al., 2011). Hence, I controlled for the effect of individual differences in my experiment by measuring some of these key dispositions, which allows stronger inferences in attributing differences in performance between the training and control group to the training itself by controlling for pre-existing participant characteristics. Furthermore, measuring these individual differences allowed for a qualification of how individual differences affect transfer success based on the specific learning materials used, which may inform future research on personalized education. The following section reviews one prominent framework for how individual differences play into heuristics and biases by highlighting two established constructs—rationality and intelligence—and discusses a third exploratory measure—recognition memory.

Individual Differences in Heuristics and Biases

In *The Rationality Quotient: Toward a Test of Rational Thinking*, Stanovich et al. (2016) detail how typical tests of intelligence (e.g., IQ tests) can fail to match up with what people normally mean by 'intelligence' because they fail to capture key cognitive features involved in decision-making and judgments, particularly those involving heuristics and biases.[4] Thus, they propose a more comprehensive assessment of intelligence that can account for the quality of people's judgments under the term 'rationality'. According to their model, what distinguishes rationality from intelligence is that intelligence only captures cognitive ability, while rationality includes both cognitive ability and thinking dispositions (Stanovich et al., 2016, p. 27). Before discussing this distinction further, it is worth discussing the model in its entirety as it applies to judgments and decision-making.

Firstly, Stanovich et al. (2016) propose a tripartite model that is an extension of the dual-process theory of judgments using systems 1 and 2 processing (Kahneman, 2011). System 1 processing generally refers to automatic, unconscious, and intuitive thinking. For example, two instances of system 1 processing described by Kahneman (2011) are the way we automatically "detect hostility in a voice" (p. 21) or answer the question "2 + 2 [= ?]" (p. 21). On the other hand, system 2 processing generally refers to slow, conscious, and effortful thinking. For example, two instances of system 2 processing described by Kahneman (2011) are when we "check the validity of a logical argument" (p. 22) or "compute the product of 17 x 24" (p. 23). In contrast, Stanovich et al. (2016) split system 2 processing into: (1) the algorithmic mind—measurable by individual differences in cognitive ability or IQ—and (2) the reflective mind—

[4] As Stanovich (2016) defines it, "the term *biases* refers to the systematic errors that people make in choosing actions and in estimating probabilities, and the term *heuristic* refers to *why* people often make these errors" (p. 26). Typically, heuristics refer to the particular mental shortcuts identified by Tversky and Kahneman's research: representativeness, availability and anchoring (Gigerenzer, 1996).

measurable by individual differences in rational thinking dispositions (Stanovich & Toplak, 2016c, pp. 22–28). Thus, based on their tripartite model, they identified three interrelated kinds of errors whereby people may fail to make rational judgments: (1) inappropriate or missing mindware; (2) a lack of conflict detection; and (3) an override error (Stanovich et al., 2016, pp. 39–62). As defined by Stanovich (2018), in the tripartite model, mindware refers to the "knowledge bases, rules, procedures and strategies" (p. 7) that a person knows and an error in inappropriate or missing mindware may occur when a person lacks the normative knowledge to make a 'correct' judgment.[5] A lack of conflict detection is related to the reflective mind and refers to when a person does not detect a conflict between their intuitive judgments and their mindware (Stanovich, 2018). An override error is related to the algorithmic mind and refers to when a person fails to override their intuition and simulate (imagine) alternative responses (Stanovich, 2018).

These three ways are interrelated because—in the case of heuristics and biases—before cognitive ability can be relevant to making rational judgments, through the reflective mind, there must first be a detection of a conflict between an intuitive incorrect response with what one has learned—their mindware (Stanovich et al., 2016, pp. 44–49). However, for conflict detection to occur, the relevant mindware must also be so overlearned and automized that it becomes an integral part of a person's automatic responses; if it is not automized, then it is only accessed by system 2 processing in computing the correct response (Stanovich et al., 2016, pp. 45–46). Thus, the three ways are interrelated because the effectiveness of cognitive ability depends on conflict detection, which in turn depends on how well mindware is instantiated. This interconnectedness

[5] While these concepts are referred to as mindware in the tripartite model, they may be referred to as semantic memory, roughly defined as memory for facts (Tulving, 2002), in the memory literature.

is critical. According to the tripartite model, if a person does not have the appropriate mindware or if it is not well-learned, then it would be inappropriate to classify biased responses as either a lack of conflict detection or override failure (Stanovich, 2018). Furthermore, the degree to which a person's mindware is well-learned and the amount of knowledge that may be required for a particular problem exist on a continuum and vary dramatically (Stanovich, 2018; Stanovich et al., 2016). Therefore, in the tripartite model, to understand what kind of error a person makes on a particular heuristics and biases problem, it is critical to consider: (1) the mindware required for a problem; and (2) the person's mindware, whether they detected a conflict between their intuition and mindware, and their ability to override their intuitions and compute the correct response. Lastly, they argue that once the mindware becomes overlearned and automatized, it should become a part of system 1 processing and may even "automatically trump [heuristics] from System 1 without needing to invoke a taxing [System] 2 override procedure" (Stanovich et al., 2016, p. 45).

Applying the Tripartite Model to the Linda Problem

According to the tripartite model to the Linda problem, when a person does not possess the normative knowledge of probability related to the likelihoods of conjunctions and their constituents, then a conjunction error should be interpreted as a mismatch in mindware. If participants have learned the required knowledge, then errors should be primarily attributed to either a lack of conflict detection or an override failure. However, in this case and in the tripartite model, gaps in a person's mindware cannot be ruled out because the mindware may not be sufficiently instantiated for conflict detection to occur. Lastly, if participants have learned the required knowledge to a higher degree and conflict detection has occurred, then errors should be interpreted as an inability to inhibit their intuitions and compute alternative responses, which is

reliant on their cognitive ability. I elaborate on each part of the tripartite model and how it will be accounted for in the proposed experiment in the following sections.

Rationality: Thinking Dispositions

Stanovich et al. (2016) operationalize the term rationality as a more complete assessment of intelligence than what IQ tests typically assess by also assessing thinking dispositions. In their book, they cite a long line of research that has consistently found that in many judgments involving heuristics and biases "rational thinking dispositions will predict variance after the effects of general intelligence have been controlled [for]" (Stanovich & Toplak, 2016c, p. 27). For instance, Toplak et al. (2011) measured a variety of individual differences that they predicted would be related to differential responses on a battery of heuristics and biases judgments. Of their measures, they included both a performance-based measure of thinking dispositions—the cognitive reflection test (CRT) —and a measure of IQ—using subtests of the Wechsler Adult Intelligence Scale. The CRT is a series of questions that are thought to provoke rapid intuitive responses and that can be answered correctly when participants use slow unintuitive thinking, when they reflect on their answers (Frederick, 2005; cf. Szaszi et al., 2017).[6] Hence, Toplak et al. (2011) interpreted higher scores on the CRT to indicate a disposition to reflect more on one's responses. Importantly, using hierarchical linear regression, they found that CRT scores were the strongest correlates of performance on the battery and explained the most variance in performance even above IQ (Toplak et al., 2011, Table 2 and 3). Thus, given that reflectivity is a

[6] Szaszi et al. (2017) provide empirical evidence to dispute the interpretation of CRT as a measure of only dispositional reflectivity. Instead, they argue that performance may reflect both reflectivity and numeracy. In their study they found that participants who reflected on their responses did not necessarily arrive at the correct responses; and most of the time, when correct responses were given, participants had already begun with the correct response. (Participants were recorded while they solved the CRT using a thinking aloud procedure.) Nevertheless, given that the CRT is one of the strongest predictors of performance on judgments involving heuristics and biases, it still serves its purpose by controlling for dispositions when estimating the effect of transfer success.

strong predictor—based on the findings of Toplak et al. (2011)—and how it is thought to capture individual differences in people's thinking dispositions—a key part of the tripartite model proposed by Stanovich et al. (2016) and a proxy for conflict detection—I used an updated version of the CRT, the CRT-7 (Toplak et al., 2013), to strengthen and qualify any inferences regarding transfer success.

Rationality: Cognitive Ability

According to the tripartite model, conflict detection by the reflective mind is a pre-requisite for the effective override of intuitive responses and effortful simulation/computation by the algorithmic mind (Stanovich et al., 2016). Hence, given that individual differences in the algorithmic mind are defined as differences in cognitive ability (Stanovich et al., 2016, fig. 2.2) and that IQ tests aim to measure cognitive ability, it follows that one may operationally use IQ scores as a proxy for how well someone is able to override their intuitive responses and perform effortful simulation/computation. Thus, to further strengthen and qualify any inferences regarding transfer success, I measured IQ using a 2 minute verbal-numeric reasoning task that was used for and provided by the UK Biobank (Lyall et al., 2016). Importantly, if the tripartite model proposed by Stanovich et al. (2016) is true, then it should be the case that when participants have learned the normative response and are told to apply their learning to the Linda problem, IQ will be the strongest predictor of rational performance.

Exploring The Importance of Individual Differences in Recognition Ability to Transfer

Presumably, increasing the number of intervening tasks between training and the Linda problem helps to further mask the purpose of the training material from trained participants, which improves the test of transfer success in my experiment under the terms of dual process theory. Hence, given that established individual differences related to performance on heuristics

and biases tasks were already controlled for, I took the opportunity to explore other potentially relevant measures. Specifically, I attempted to explore individual differences in how easily mindware becomes 'automatized', such that it becomes so memorized and practiced that it becomes part of a person's automatic and intuitive responses—a part of their system 1 (Stanovich & Toplak, 2016a, fig. 3.1). In other words, what are the key individual differences in predicting how quickly someone learns and how well they transfer their learning? To my knowledge, individual differences in learning and transfer are an underexplored component of the tripartite model and transfer research in general (e.g. Nichols et al., 2021; Owen et al., 2010; Simons et al., 2016), despite being considered an important part of the phenomena of transfer (Barnett & Ceci, 2002). Historically, IQ has been consistently identified as an important positive predictor of successful learning and transfer (Barnett & Ceci, 2002; e.g. Evans et al., 2010). Hence, I explored a different candidate for individual differences in learning and transfer: recognition memory.

Here, recognition memory refers to how well a person correctly recognizes the match or mismatch between a currently presented and previously learned stimuli. Those who are more able to correctly recognize what was previously presented are operationally defined to have better recognition memory and hence be better in: (1) learning; and/or (2) remembering information. All other things being equal, those who are better in learning will have more of the training material 'stored in their memory'. All other things being equal, those who are better at remembering will be better able to retrieve what they have 'stored in their memory'. Thus, I speculate that—all other things being equal—those with better recognition memory may be better at recognizing the similarities between what they learned and a new problem, which would presumably facilitate transfer of learning. Furthermore, some judgment theorists have framed

recognition as one of the key heuristics in adaptive judgments (Gigerenzer & Gaissmaier, 2011; Goldstein & Gigerenzer, 2002). Given that recognition may be important for transfer and how it is theorized to be a core heuristic, measuring individual differences in recognition seemed promising.

To measure recognition memory, I used a classic verbal paired associates (VPA) recognition task, which tests how well a person correctly recognizes previously learned word pairs when presented among lure word pairs. Lure word pairs comprise previously shown words that have been recombined. Presumably, those who are better at correctly identifying the previously shown word pairs are generally better able to recognize the match and mismatch between what is currently being presented and something that they have previously seen. Hence, I predicted that—controlling for thinking dispositions and cognitive ability—those with better recognition memory will be better at recognizing the similarities between their training material and the Linda problem, which will lead to higher transfer success. Furthermore, given that the VPA is simple to administer and is widely used (e.g. I. A. Clark et al., 2018; Paivio, 1965)—even appearing in the Wechsler Memory Scale (Kent, 2017)—it appeared to be the most appropriate measure of recognition memory for the purposes of this study. Thus, I incorporated it as my third measure of individual differences.

Creating New Learning Materials to Boost Transfer Success

Lastly, there is research suggesting that abstract (non-physical) learning materials are superior to their concrete (physical and highly salient) counterparts (Kaminski et al., 2008, 2013; c.f. Bock et al., 2011; J. M. Clark & Paivio, 1991; Jones, 2009b, 2009a). Given that the learning materials used by Agnoli and Krantz (1989) are highly concrete through their use of Venn diagrams and salient real world examples, I adapted the learning materials of Agnoli and Krantz

(1989) to create a set of matched learning materials that only differed in the presence or absence of diagrams and textual references to diagrams (see Appendix B for learning materials). Creating matched learning materials without diagrams is important because it allows one to compare matched abstract vs. concrete learning material, that are purely textual, in the future. (I found it overly difficult to create abstract learning materials with diagrams.) Furthermore, given the finding by Agnoli and Krantz (1989) that trained participants answered only about 56% of the Linda and Linda-like problems correctly on-average (see Table 3 of their paper), it was important to determine if performance could be improved further, and the use of abstract learning materials was one potential avenue.

Objectives and Hypotheses

Based on the reviewed literature, this experiment was designed to inform our understanding of whether education can inhibit heuristics by: (1A) testing if transfer success occurs when the relationship between the learning material and Linda problem is obscured; (1B) qualifying claims of transfer success by accounting for dispositions that are established to predict performance on heuristics and biases tasks; (2) modelling how established individual differences and recognition memory performance (an exploratory measure) relate to transfer success and performance on the Linda problem; and (3) comparing how learning materials with and without diagrams differentially affect transfer and whether any differential effects interact with individual differences. The following sections will restate each objective and outline key testable hypotheses with figures of theoretically informed predictions. Importantly, although the visualizations of the predictions in this section imply that I have precise point-estimates in mind, most of the point values were chosen to best express the predictions of the relations between

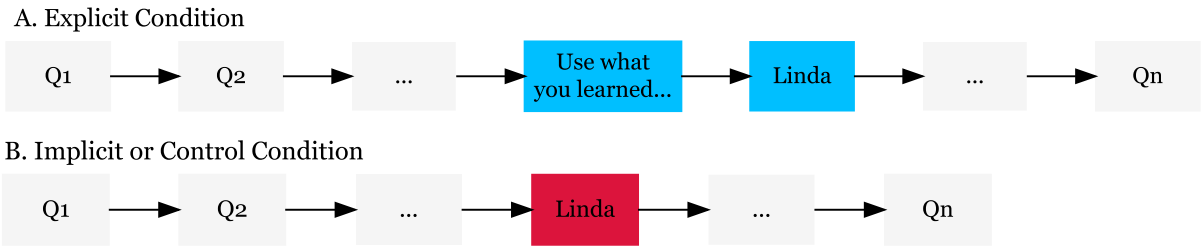
different variables made by the associated theory, not some precise value based on prior literature.

Objective 1: Does Transfer Occur When the Relevance of the Learning Material to the Linda Problem is Obscured?

To address the main question of whether a true transfer success will occur when participants are not cued to the relationship between the learning material and Linda problem, participants were randomly assigned to one of three learning-instruction conditions: (1) a control group that undergoes no training; (2) an explicit group that undergoes training and receives both a prompt before training that what they learn will be tested later, and another prompt, immediately before they see the Linda problem, to apply what they learned; and (3) an implicit group that undergoes training, but receives no such prompts. These three learning conditions are henceforth referred to as the three between-subjects levels of the 'prompt' condition. To further mask the relationship between the learning material and the Linda problem, there were several intervening tasks between the learning and the Linda problem (see Figure 5); and the Linda problem was embedded in a battery of heuristics and biases questions unrelated to the training (see Figure 1), taken from Toplak et al. (2011). For a discussion of why only the Linda problem was used and why I argue that there should only be one conjunction problem in this experiment, see Appendix A.

Figure 1

Variations in How The Linda Problem was Embedded in The Heuristics and Biases Battery (Toplak et al., 2011) Based on Learning Condition



Note. Each rectangle represents a different questions taken from the heuristics and biases battery described by Toplak et al. (2011).

Primary Confirmatory Hypotheses

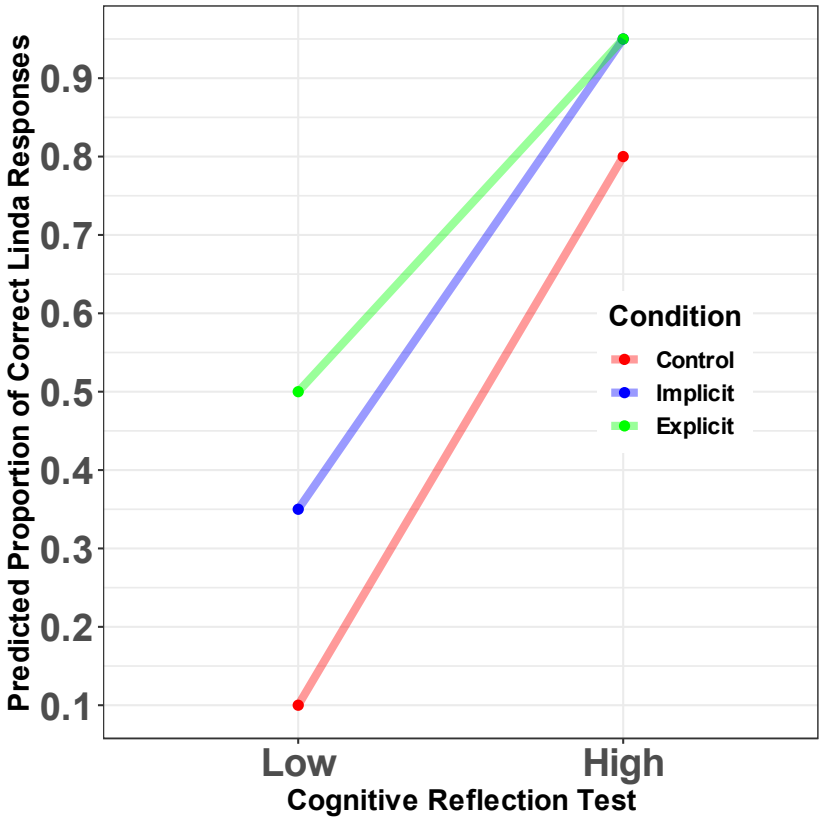
Figure 2 illustrates the main predicted effects for the first objective. I hypothesized that if the training described by Agnoli and Krantz (1989) inhibited the representative heuristic, then at least 15% more participants in the implicit condition will perform better than the control group.[7] Based on the findings reported by Agnoli and Krantz (1989), I had prior reason to believe that the training would likely produce some improvement in trained participants compared to the control. However, the primary question I chose to address with this experiment is whether the improvement without explicit prompts could be so small that it could be considered negligible, not to 'prove' the null significance hypothesis. As many others have said, the null (significance) hypothesis can nearly always be assumed to be false (Cohen, 1994; Meehl, 1978) and with a large enough sample size, a statistically significant difference will almost always be found (Meehl, 1990). Hence, this 15% criterion was chosen because any amount lower than 15% is arbitrarily presumed to be practically inconsequential, especially when one considers the

[7] Throughout this paper, I refer to the difference in the proportion of correct responses between conditions in terms of percentages for simplicity. For example, a 15% difference between control and implicit when 30% of the control makes conjunction errors means that either 15% or 45% of the implicit group made conjunction errors. It does not mean that either 25% or 35% of the implicit group made conjunction errors.

possibility of forgetting over time. In my literature review, I did not find substantial discussion of minimum meaningful effect sizes in this context, but one was needed to guide power analyses and equivalence testing, both of which were necessary. Thus, I assumed that a 20% difference is substantial and that a 10% difference is unsubstantial, so I pegged the minimum meaningful effect size at 15%.

Figure 2

Anticipated Proportion of Correct Linda Responses Based on Learning Condition and CRT Scores



Note. CRT scores are split into low and high to facilitate interpretation. Higher proportion of correct Linda responses is better.

If at least a 15% improvement is found, then I will interpret this to mean that education truly can inhibit heuristics, even when people face the Linda problem without any prior prompts

to apply their training—consistent with prior claims in the literature (Agnoli, 1991; Agnoli & Krantz, 1989; Kahneman, 2003). For example, the titles of Agnoli and Krantz (1989) and Agnoli (1991) are (1) *Suppressing Natural Heuristics by Formal Instruction: The Case of the Conjunction Fallacy* and (2) *Development of Judgmental Heuristics and Logical Reasoning: Training Counteracts the Representativeness Heuristic*, respectively. Also, Kahneman (2003) writes:

> Agnoli and Krantz (1989) reported that brief training in the logic of sets improved performance in a simple version of the Linda problem. The findings indicate that the accessibility of statistical heuristics can be enhanced in at least three ways: by increasing the vigilance of the monitoring activities, by providing stronger cues to the relevant rules, and by extensive training in applied statistical reasoning. (pg. 711)

In context, the implication of this quote is that extensive statistical training can inhibit the representativeness heuristic *independently* of increasing monitoring and providing stronger cues to apply learning.[8] Regardless of whether this implication was intended or not, to my knowledge, there has been no prior study that has critically examined whether statistical training truly can inhibit inappropriate heuristics independently of cuing and attempts to increase 'monitoring'. Hence, this experiment's primary purpose was to clarify whether statistical training—by-itself and all other things being equal—can inhibit heuristics.

If I fail to find that the implicit and control conditions are significantly different by at least 15%, then I will conclude that contrary to prior literature, it has not yet been shown that education can inhibit the heuristics elicited by the Linda problem on dual process theory's terms.

[8] To my knowledge, Agnoli and Krantz (1989) use the same Linda problem that was first reported by Tversky and Kahneman (1983).

It also highlights that other educational studies that instruct learners to apply what they learn to transfer problems should strongly consider evaluating the role of cues to apply learning (Jones, 2009b; e.g. Fyfe et al., 2015; Kaminski et al., 2008; McNeil & Fyfe, 2012; Sedlmeier & Gigerenzer, 2001). Given that the CRT is established to be one of the strongest predictors of performance on heuristics and biases tasks (Toplak et al., 2011), I will also model the CRT scores with the training conditions to qualify the claim that transfer effects were due to training and not a difference in the distribution of CRT scores between conditions.

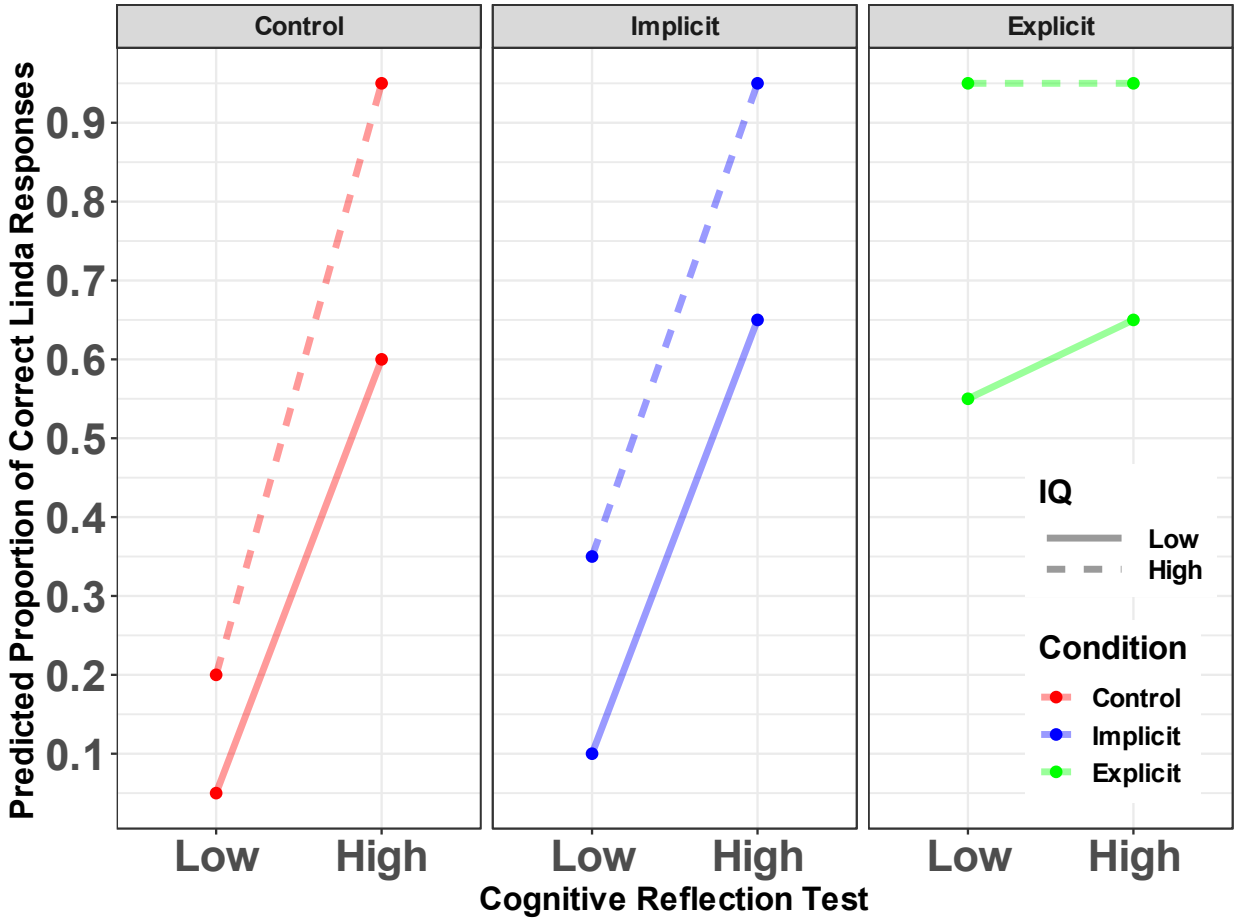
Objective 2: How do Individual Differences Relate to Transfer Success and Performance on the Linda Problem?

Confirmatory Hypotheses Based on Established Individual Differences

Based on the tripartite model, Figure 3 illustrates my predictions for how IQ and thinking dispositions (CRT) will predict performance in each learning condition. Given that conflict detection is thought to be a critical pre-requisite to performance on heuristics and biases tasks (Stanovich & Toplak, 2016b), and replicating Toplak et al., (2011), I hypothesized that higher CRT scores will predict better performance on the Linda problem. Based on how cognitive ability is thought to be required for correct responses following conflict detection (Stanovich & Toplak, 2016b), I predicted that there will be an interaction between IQ and CRT, such that although higher IQ scorers will generally perform better than lower IQ scorers, this difference is smaller when both score lower in CRT, and larger when both score higher in CRT. Furthermore, I predicted that because the explicit prompt condition removes the need for conflict detection, then CRT will no longer be an important predictor of task performance in the explicit learning condition, such that high IQ scorers perform at ceiling and lower IQ scorers will perform better than in the control condition.

Figure 3

Anticipated Proportion of Correct Linda Responses Based on CRT and IQ Scores Facetted by Learning Condition



Note. CRT and IQ Scores are split into low and high—and effects are exaggerated to facilitate interpretation. Higher proportion of correct Linda responses is better. Points were largely chosen based on the qualitative predictions of the tripartite model proposed by Stanovich et al. (2016) and not on the precise numerical values. Predictions are also made under the assumption that CRT and IQ scores are perfectly separable into reflectivity and cognitive ability, respectively.

If these predictions are correct, then my findings will support the differential importance of thinking dispositions and cognitive ability, consistent with the framework proposed by Stanovich and Toplak (2016). However, if the predictions are not supported, then given the strong prior for each of the hypotheses, this would suggest either a methodological flaw or an

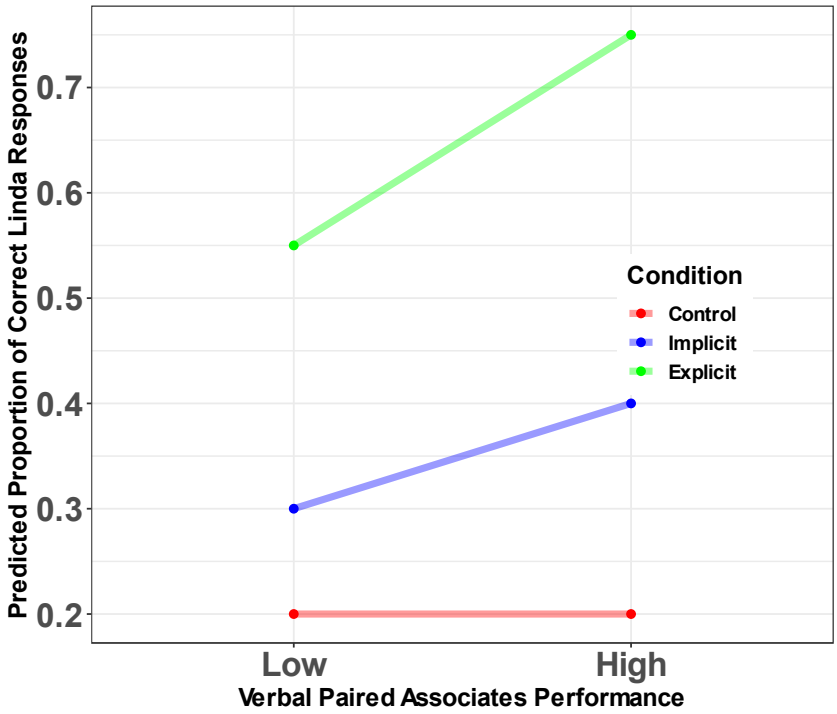
underappreciated boundary condition on the validity of their framework. Given that CRT scores are highly correlated with cognitive-ability/IQ (Toplak et al., 2011), my initial prediction that I would observe a differential importance of thinking dispositions and cognitive ability may have been unwarranted.

Exploratory Hypothesis

Figure 4 illustrates my prediction for how individual differences in recognition memory relate to transfer success in the proposed experiment. Controlling for all other key covariates and assuming that remembering the learning material is difficult enough for individual differences in recognition memory in a healthy population to measurably affect transfer success, I predicted that trained participants with better recognition memory will generally perform better—all other things being equal—particularly in the explicit learning condition. I also predicted that there would be virtually no effect of individual differences in recognition memory in control group participants, when controlling for all other important covariates.

Figure 4

Anticipated Proportion of Correct Linda Responses Based on Learning Condition and Individual Differences in Verbal Paired Associates Recognition Memory Performance



Note. Controlling for CRT and IQ.

Objective 3: Do Learning Materials With Diagrams Improve Transfer More Than Without on Inhibition of Biased Judgments? Do These Effects Interact With Individual Differences?

To preface this subsection, much of the substance behind this objective was conceived after the data were collected and analyzed. However, it was always an unwritten goal that the purely textual learning materials be tested, so that a future experiment could compare abstract and concrete textual learning materials. As aforementioned, I thought it would not be possible to make learning materials that were both abstract and contained visual diagrams, which led to the current roundabout approach.

To test this third objective regarding learning materials with and without images, half of the participants from each of the previously mentioned training groups—implicit and explicit—were further randomly given learning materials with images, and half were given learning materials without images, as shown in Figure 5. To create the learning material with images, the learning materials provided by Agnoli and Krantz (1989) were adapted, primarily through the use of new diagrams and a more logically consistent design language created using the R package *eulerr* (Larsson & Gustafsson, 2018). The learning material without images was created by removing the diagrams and any textual references made to them. There were no strong prior hypotheses regarding the differences between the presence and absence of visualizations on the proportion of participants that would make conjunction errors on the Linda problem, nor how individual differences may interact with learning materials.

Method

Participants

There were a total of 554 unique participants who signed the consent form in this experiment, 205 from the local York University Undergraduate Research Participant Pool (URPP) and 349 from the online platform Prolific (*Prolific*, 2022). (Uniqueness was determined based on their identification number from either URPP or Prolific.) The experiment was conducted remotely with minimal experimenter interaction using both PsychoJS (Peirce et al., 2019)—hosted through Pavlovia (Bridges et al., 2020)—and Qualtrics (*Qualtrics*, 2022). Specifically, after I posted the experiment, a set batch of participants were able to freely sign-up and independently begin the experiment at any time. Prolific participants were free to message me anonymously using Prolific's built-in messaging platform and URPP participants could send me emails.

To participate in this study, participants must have been registered in either URPP or Prolific as being between 18 and 25 years old; speaking English as their primary language; and having normal or corrected-to-normal vision and adequate hearing. The posting also indicated that they must participate using either a desktop or laptop computer with a keyboard and mouse. Prolific participants from the USA, UK, Australia and Canada who met the above criteria were free to participate in this study, so long as they had an approval rating of at least 80% and participated in between 5 and 10000 other Prolific studies. Approval ratings refer to the proportion of posted experiments that a Prolific participant was previously compensated for.

Regarding compensation, so long as URPP participants signed the consent form and did not revoke their consent, their course credit would be granted—regardless of data quality and completeness—and they were also free to complete alternative written assignments if they did not want to participate in any experiments whatsoever. Similarly, any Prolific participants who signed our consent form were compensated 9.50 US Dollars.

Regarding which participants data were used for the forthcoming analyses, if a participant produced clearly questionable data quality their data were removed from all analyses, leaving 453 participants of the original 554. The allocations of the remaining participants are shown in Table 1 and the associated self-reported demographics are shown in Table 2. Here, clearly questionable data quality was defined as when a participant's data: (1) was incomplete; (2) contained uncorrectable nonsensical responses on any short answer question without further explanation; (3) provided experiment feedback explicitly stating that they did not pay attention to the experiment; (4) was marked by Qualtrics as 'Spam'; (5) indicated that they revoked their consent; (6) indicated that they took more than 180 minutes to complete the entire study from the moment they submitted their consent form; or (7) contained clear indication or communication

that there was some technical difficulty that could threaten the validity of the experiment—e.g. repeating portions of the experiment.

Sensitivity Dataset

Given concerns regarding the quality of online data collection, I created a 'sensitivity dataset' based on a subset of the baseline 453 participants dataset, for which more stringent exclusion criteria were applied. Control group participants must have answered more than 75% of the reading comprehension questions correctly and trained participants must have answered both of the questions in their learning materials correctly. Also, all sensitivity participants must: (1) have had no prior knowledge of the Linda problem or Daniel Kahneman and Amos Tversky's research; and (2) have not been classified as multivariate outliers, based on their Mahalanobis distances at a $p \geq 0.001$ threshold. A list of measures used to compute Mahalanobis distances can be found in Appendix C.

Table 1

Participant Counts per Prompt and Learning Material Condition: Baseline (Sensitivity)

Prompt	Learning Material		
	Control	Graphs & Text	Text
Control	93 (38)		
Implicit		96 (68)	88 (63)
Explicit		86 (58)	90 (53)

Note. Prompt refers to whether the participant was in: (1) the control group with no training; (2) the implicit group, where they received no explicit indication to apply their training or what the purpose of it was; and (3) the explicit group, where they received an explicit prompt before the training that it would be relevant to a latter problem and another prompt directly before the Linda problem to apply their training. Learning material refers to whether the use of graphs & text—meaning images with text—or only text were involved. Participants in the control group received no training, hence they belong in a distinct cell.

Unbracketted numbers represent the baseline 453 participants used throughout the reported results. Bracketted numbers represent the subset of 280 participants selected that passed more stringent data quality measures and that were used to test the robustness of initial results.

Table 2

Participant Demographics per Prompt and Learning Material Condition

Prompt	Learning Materials	Mean Age (SD)	Gender					
			Gender fluid	Man	Non-binary	Prefer not to say	Transgen-der man	Woman
Control	Control	20.58 (1.74)		43	3	1		46
Explicit	Graphs & Text	21.71 (2.04)		44	4			38
	Text	21.93 (2.25)		33	5	3		49
Implicit	Graphs & Text	21.74 (2.17)		35	5			56
	Text	21.55 (2.23)	1	45	2		1	39

Note. Due to the specifics of the ethics protocol, only the self-reported gender of every participant was collected in this experiment. One Prolific participant did not report their birthday. No participants reported themselves as transgender woman.

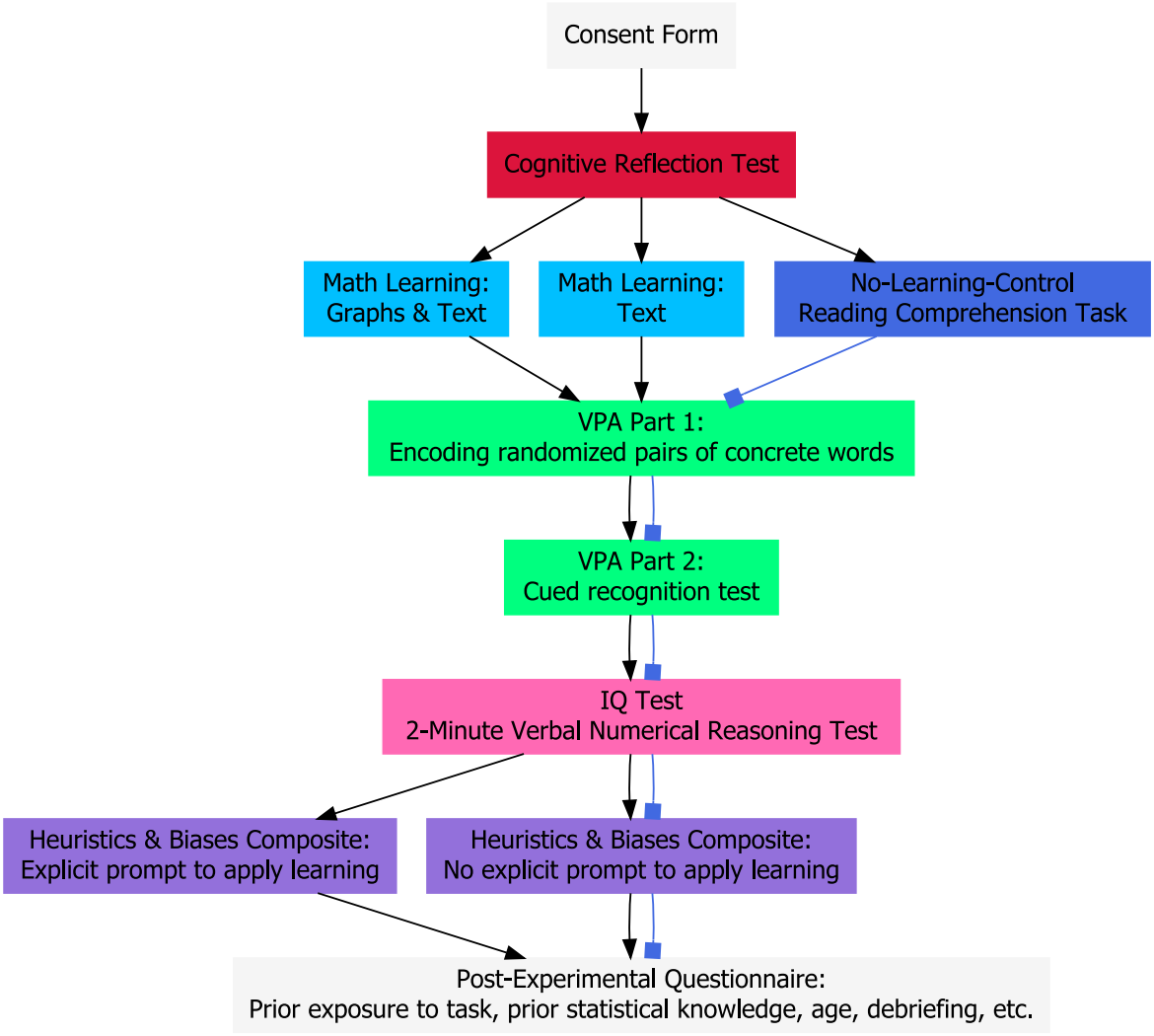
Procedure and Tasks

Shown in Figure 5 is a flowchart of the experimental procedure. Each level moving vertically represents the flow over time and each horizontal branch represents where subjects were randomly assigned to their between-subject conditions for: (1) the kind of math learning—graphs & text or text-only, or control—and (2) whether they received an explicit prompt to apply their learning to the Linda problem or not prior to and during the heuristics and biases composite. Note that the control group never received a prompt to apply learning, a pathway that is shown in Figure 5 through a distinct set of royal blue arrows. Each task will be discussed in the following subsections. In total, the experiment was estimated to take about 55.7 minutes and the posting for

the experiment advertised that it should take about an hour to complete. The completion time for the subset of 453 participants is shown in Figure 6 and in Table 3. Here, completion time refers to the time elapsed from the moment participants completed the consent form and load the first part of the experiment to the time they submitted the post-experimental questionnaire.

Figure 5

Experiment Procedure in Flowchart Form



Note. Dark blue lines show how control group participants never receive a prompt to apply their learning.

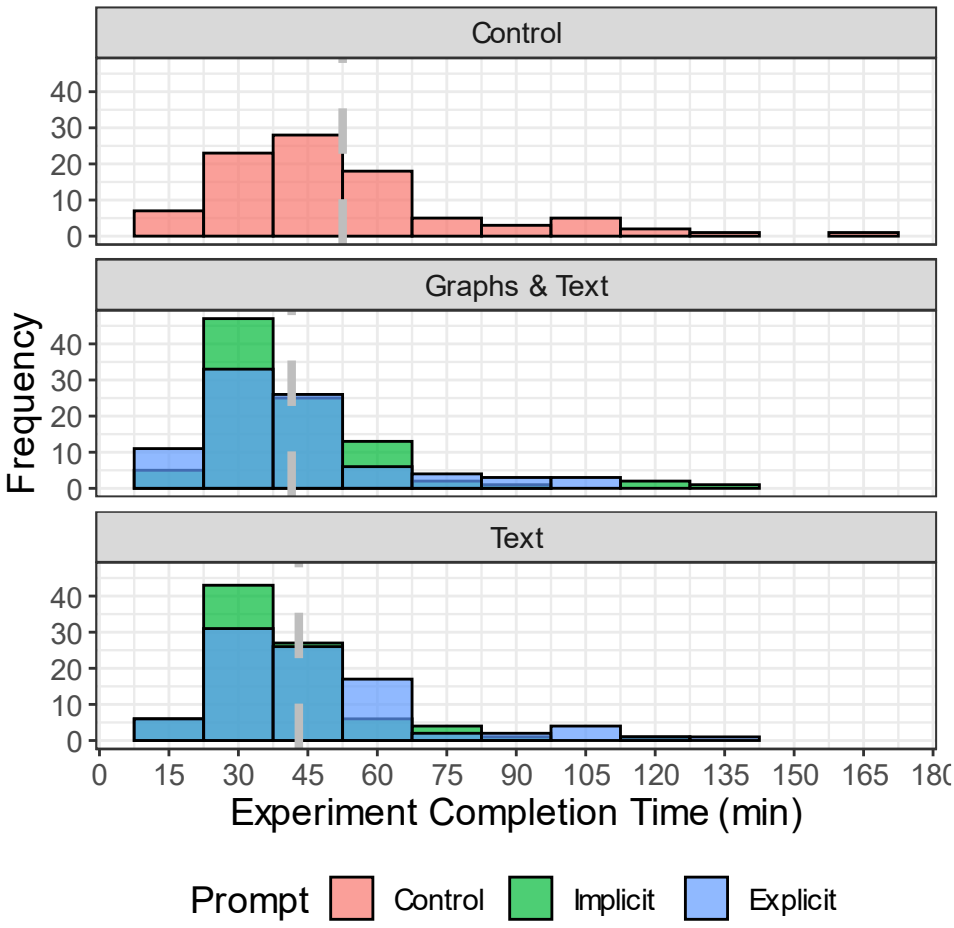
Table 3

Experiment Completion Time Statistics by Prompt and Learning Material for 453 Participants

Prompt	Learning Materials	Experiment Completion Time in Minutes					
		M (SD)	Minimum	25%ile	50%ile	75%ile	Maximum
Control	Control	52.51 (28.12)	14.38	32.63	47.00	61.42	167.70
Implicit	Graphs & Text	41.59 (20.59)	16.88	28.36	35.03	49.61	136.40
Implicit	Text	39.36 (16.9)	15.28	29.17	33.31	45.48	120.88
Explicit	Graphs & Text	41.4 (21.05)	15.95	27.57	36.48	45.96	109.98
Explicit	Text	46.66 (23.4)	8.80	30.93	40.37	55.72	127.53

Figure 6

Histograms of Experiment Completion Times for 453 Participants Under 180 Minutes



Note. The gray dashed line represents the mean completion time for each learning material condition computed on the set of participants—453 participants—who completed the experiment in under 180 minutes.

Measuring Thinking Disposition: Cognitive Reflection Test

To measure participants' predisposition to reflect on their answers, I used the CRT-7 provided by Toplak et al. (2013). As aforementioned, the CRT-7 is a set of 7 questions each with an intuitive incorrect and non-intuitive correct answer. Questions were shown one-at-a-time in the same order across all participants. Given 7 questions and 1 minute per question, this task was estimated to take 7 minutes.

Learning Material

The learning material used to train participants for the Linda problem was adapted from Agnoli and Krantz (1989) and is provided in Appendix B. To fulfill objective 1, participants in the explicit condition were prompted with a message indicating that the coming learning materials would be relevant to an upcoming question in the experiment, whereas participants in the implicit condition received no such message. To fulfill objective 3, a pair of matched learning materials was made where one incorporated Venn diagrams to aid participants' understanding of the material and the other did not (see Appendix B). Based on Agnoli and Krantz (1989), it was estimated that participants would take around 20 minutes to learn the material and answer its in-lesson questions. Participants in the control group completed a comparably-long 20-minute reading comprehension control task. Following Agnoli and Krantz (1989), there were two multiple choice questions to assess their understanding of the material within the learning material.

Filler Task.

Instead of reading learning material, control group participants worked through a comparably-long 10th grade reading comprehension task provided by CommonLit on the Roaring Twenties (Kubic, 2016).

Measuring Recognition Memory: Verbal Paired Associates

Individual differences in recognition memory were measured using a VPA task. VPA is a longstanding memory testing paradigm (e.g. I. A. Clark et al., 2018; Paivio, 1965) of how well people remember pairs of words. Here, it was operationalized as a proxy for individual differences in recognition memory based on participants' correct or incorrect recognition of whether a word pair presented at test was previously shown or not. Given the complexity of the

measure, I explain it in two parts—first, a conceptual version, then a precisely specified version. My version of the VPA was programmed using PsychoJS—the online version of PsychoPy (Peirce et al., 2019)—and hosted on Pavlovia (Bridges et al., 2020).

Conceptually speaking, the task is split into two main parts, an encoding phase and a recognition phase (see Figure 5). In the encoding (study) phase, randomized pairs of words are shown to the participant one pair at a time and participants are asked to try to remember as many pairs of words as possible for a later test. In the recognition (test) phase, half of the exact pairings of words that were previously shown in the encoding phase are presented again among recombined word pairs (lures), and participants are asked to judge whether they saw each word pair previously or not—i.e., whether they recognize it. Also, each of the individual words in the other half of the previously shown word pairs are recombined to create completely new word pairs (lures), which are then randomly shuffled with the aforementioned exact word pairs in the recognition phase. By testing both recombined and previously shown word pairs, the task is made more difficult and accounts for the scenario where a participant responds that all the tested word pairs were previously shown. Hence, the first measure of this task is the hit rate—the proportion of previously shown word pairs that participants correctly recognized. The second measure of this task is based on the recombined word pairs, the false alarm or false positive rate—the proportion of novel word pairs that participants incorrectly judged as having been previously shown before. The false alarm and hit rate can be combined to create a measure called discrimanibility, which balances the two into a single measure of memory ability (Snodgrass & Corwin, 1988; e.g. Ward et al., 2020).

In detail, participants complete the VPA twice, once in a 6 trial practice and a longer 30 trial main task, which was used to measure their recognition memory. Lists of word pairs were

created by randomly selecting from a list of words of highly imageable concrete objects kindly provided by I. A. Clark et al. (2018). Beginning with the encoding phase, participants see pairs of words to be remembered one-at-a-time—one word on the top half of the screen and another on the bottom half—for 2.5 seconds each , with word pairs separated by an intervening 1-second fixation cross. Then, before the recognition phase, the program: (1) shuffles the order of the previously shown word pairs; (2) splits this shuffled list into two parts called true old and false old; (3) recombines false old word pairs by randomly changing what word is paired with what, such that that recombined word pairs are never identical to the encoding phase; and (4) combining this set of false old recombined word pairs and true old word pairs before shuffling one last time. Importantly, in the recognition phase, regardless of whether the word pair is true old or false old, words that were previously shown on the top half of the screen will still be shown on the top half and those shown on the bottom half shown on the bottom, so that recognition should be based on the exact pairing of the words and not memory of the spatial location of a given word. Assuming that each encoding trial takes 3.5 seconds and that each recognition trial takes 6 seconds (1-second fixation cross and 5 seconds to make a judgment, as-per I. A. Clark et al. (2018)); that there are 6 practice trials for familiarizing participants with the task and 30 test trials; and that reading the instructions will take 1 minute this task was estimated to take about 6.7 minutes.

Measuring General Cognitive Ability: 2 Minute UK Biobank Form

To measure participants' cognitive ability as a proxy of their ability to perform cognitive decoupling and effortful calculation, I used a 2 minute cognitive ability test that was used for and provided by the UK Biobank (Hagenaars et al., 2016; Lyall et al., 2016). This task comprises 13 questions that were individually primarily verbal or numerical in nature. Using Qualtrics, all 13

questions were presented on the same page in the same order for all participants, with a timer indicating the remaining time for each participant to submit their responses. Before seeing this section, participants were told that they would have 2 minutes to complete as many questions as quickly and accurately as possible. Participants automatically advanced to the next section after 2 minutes and their page was disabled from going back inorder to prevent them from retaking this portion of the survey.

Heuristic and Biases Composite Embedded with Linda Problem

In the last phase of the experiment, participants were presented with a battery of heuristics and biases questions provided by Toplak et al. (2011), which already included a slightly modified Linda problem—their version of the Linda problem changes the order of the to-be-rated options. Given that each of the outcome bias and framing problems are two-part questions that cannot be presented adjacently (M. Toplak, personal communication, February 4, 2022), they were split into the beginning and end of the battery. Otherwise, the rest of the questions in the battery—including the Linda problem—were presented in a random order for each participant, one-at-a-time. Furthermore, each question in the battery is a previously studied problem in the heuristics and biases tradition (Toplak et al., 2011). Thus, performance on non-Linda heuristics and biases problems could be used to test the assumption that the content—and hence benefits—of the learning material should be exclusively related to the Linda problem. Given 15 problems (Toplak et al., 2011) and 1 minute per problem, this task was estimated to take 15 minutes.

Post-Experimental Questionnaire

The post-experimental questionnaire was used to measure participants' self-reported justifications for their chosen rankings on the Linda problem, demographic information, prior

exposure to content used in the experiment, prior statistical training, their rating of their experience with the experiment, and for their comments on the experiment. For the justifications on the Linda problem, participants were shown the Linda problem again with their non-modifiable rankings and asked to explain their rankings, then they were shown the same content, but asked to explain why they ranked 'Linda is a bank teller and is active in the feminist movement' relative to 'Linda is a bank teller' in the way they did. In terms of prior exposure to content used in the experiment, they were asked if they had ever seen the Linda problem, heard of Daniel Kahneman and Amos Tversky's research, or seen any of the prior content used in the experiment and if so, what. Prior exposure to the CRT is a potentially important concern highlighted in prior literature because if they recognize that the questions are known to elicit rapid incorrect answers, then using the CRT as a measure of dispositional reflectivity may be invalidated (Toplak et al., 2011, 2013; cf. Bialek & Pennycook, 2018; Meyer et al., 2018).[9] This portion was estimated to take 5 minutes.

Analysis

This section discusses the general statistical techniques and analysis procedures applied for each objective. The data for this experiment were analyzed primarily through the use of various multiple binomial logistic regressions using the R programming language (R Core Team, 2021). Binomial logistic regression models the odds of one binary outcome over its complement—meaning correct to incorrect responses on the Linda problem, here—based on the selected predictors (Fox, 2016). Several regression models for each objective were fit for each question in the interest of finding the most parsimonious and well-fitting model.

[9] Bialek & Pennycook (2018) and Meyer et al. (2018) argue that repeated exposures to the CRT does not invalidate its use as a measure of reflectivity.

Regarding null hypothesis testing of the beta coefficients in a binomial logistic regression context, although the software-defined default in R uses Wald tests, Wald tests tend to be less reliable than likelihood-ratio tests (Fox, 2016, pp. 425–426), so likelihood-ratio tests were used as frequently as possible for this purpose using the R package *car* (Fox et al., 2022). When likelihood-ratio tests were infeasible, I defaulted to Wald tests of the beta coefficients.

After multiple candidate regression models were fitted to answer a specific question, I compared summary statistics of each of the models to choose the most parsimonious and best fitting model based on their Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and a likelihood-ratio test comparing a given regression model to the fully saturated model, as implemented in the R package *vcdExtra* (Friendly et al., 2022; Friendly & Meyer, 2015, p. 267).

After selecting the best regression model, I computed post-hoc tests based on either the average marginal effect (AME) through the R package *marginaleffects* (Arel-Bundock, 2022) or the marginal estimated at the mean (MEM) through *emmeans* (Lenth et al., 2022). Historically, contrasts or the visualizations of multiple regression are typically done by holding the non-specified predictors at their mean or typical values (Hanmer & Ozan Kalkan, 2013)—also known as the MEM. However, this approach has been criticized on the grounds that holding other predictors at their means may lead to an invalid effect size (Hanmer & Ozan Kalkan, 2013). Hence, the alternative is to average the effect of a contrast across all values of the other predictor variables (Hanmer & Ozan Kalkan, 2013)—also known as the AME. Here, I used AME as frequently as possible, but otherwise defaulted to the MEM for contrasts. Generally, when I state that covariates are controlled for, this refers to the use of MEM. Otherwise, when I state that the effect is averaged across other covariates, this refers to the use of AME. For visualizations of

regression lines, only MEM were used, as provided by *ggeffects* (Lüdecke, 2018). However, it is important to note that the regression visualizations only show what the model predicts and not the raw data itself. In other words, the underlying data may 'contain' an interaction, even if one is not apparent in the diagram. Arrangement of plots was done using *cowplot* (Wilke, 2020) and colors were frequently selected using *colorspace* (Zeileis et al., 2020). All tables were made using *flextable* (Gohel et al., 2022). Every flowchart in this document was made using *DiagrammeR* (Iannone, 2020). Correlation matrices in Appendix D were made using *ggally* (Schloerke et al., 2021).

The same analyses were re-run using the sensitivity dataset to support results found using the baseline dataset, while also applying regression model diagnostics programmed by *car* (Fox et al., 2022) based on examples provided by Friendly & Meyer (2015).

Special Analyses and Measures

Given that objectives 1 and 3 involves testing the absence of an effect, equivalence testing was used as the most powerful statistical technique to establish a difference of less than 15% between each condition (Lakens, 2017), using Fisher's exact z-test as implemented in the R package *TOSTER* (Lakens, 2017). In contrast to null hypothesis testing, which attempts to reject an absence of an effect, equivalence testing attempts to reject the presence of a user-specified effect (Lakens, 2017).

For the VPA task, two performance measures were computed. The first is the corrected proportion of hits—instances where a previously shown word pair is correctly recognized as being old—calculated as the number of hits + 0.5 divided by number of learned word pairs + 1, as provided by Snodgrass and Corwin (1988). This correction helps to account for extreme hit frequencies (Snodgrass & Corwin, 1988). The second is the *d*-prime measure, which combines a

participant's proportion of hits and false alarms into a single measurement and is commonly used to assess recognition memory (e.g. I. A. Clark et al., 2018; Ward et al., 2020). As computed by *psycho* (Makowski, 2018), the *d*-prime for each participant was the difference between the *z*-values of their hit and false alarm rate. Given that *d*-prime accounts for false alarms, it was used as the main measure of recognition memory.

Results

Results in this section are organized based on each objective. Summary statistics for each task and a bivariate correlation matrix of each of the key tasks in this experiment can be found in Appendix D. Unless stated otherwise, the results presented throughout are based on the baseline set of 453 participants and are consistent with the sensitivity analyses using the subset of 280 participants selected based on more stringent criteria. Furthermore, all measures of individual differences reported here are continuous because treating them as categorical frequently resulted in either: (1) a more complex model compared to being continuous; or (2) an outright failure to fit the model due to having too few values at certain combinations of individual differences.

Objective 1 Results: Does Transfer Occur When Relevance of the Learning Material to the Linda Problem is Obscured?

Comparing the Effects of Training With or Without Prompting Against a Control

As shown in Table 4, 84% of participants in the control group made conjunction errors and this proportion was lower in trained participants. Each condition of the 'prompt' variable that received training—implicit and explicit—was compared to the control group using a Wald test of the beta coefficients of a logistic regression model of Linda problem accuracy using the prompt as the sole predictor. Although Table 4 further separates each prompt condition into separate learning material groups—graphs & text or text-only—the analyses in this objective did

not differentiate based on learning material. Results of the 'Prompt' regression model (Table 5) indicated that the odds of answering the Linda problem correctly were significantly greater for each trained group compared to the control group (*OR* = 3.19, 95% CI [1.74, 6.16] for the implicit condition and *OR* = 7.00, 95% CI [3.83, 13.54] for explicit).[10] A follow-up MEM contrast found that explicit condition participants were significantly more likely to answer correctly compared to the implicit condition (*OR* = 2.19, 95% CI[1.32, 3.63], $p < 0.001$).

Table 4

Proportion of Participants in Each Condition That Made Conjunction Errors

Prompt	Learning Material		
	Control	Graphs & Text	Text
Control	0.84 (0.04)		
Implicit		0.65 (0.05)	0.59 (0.05)
Explicit		0.44 (0.05)	0.41 (0.05)

Note. Each cell shows the proportion of participants in each prompt and learning material condition that made conjunction errors on the Linda problem. Standard errors of each proportion are shown in brackets and were computed as done in Agresti (2019, p. 8).

[10] OR stands for odds-ratio.

Table 5

Objective 1 Logistic Regression Table

Variable	Prompt			Prompt & CRT-7 Additive			Prompt & CRT-7 Interaction		
	β (SE)	OR 95% CI	*p*	β (SE)	OR 95% CI	*p*	β (SE)	OR 95% CI	*p*
Control-Intercept	-1.649 (0.282)	0.192 [0.106, 0.324]	0.000	-2.312 (0.328)	0.099 [0.05, 0.183]	0.000	-2.05 (0.488)	0.129 [0.045, 0.309]	0.000
Explicit	1.946 (0.321)	7.003 [3.827, 13.536]	0.000	1.981 (0.328)	7.247 [3.903, 14.211]	0.000	1.656 (0.55)	5.24 [1.893, 16.761]	0.003
Implicit	1.161 (0.32)	3.193 [1.744, 6.162]	0.000	1.161 (0.327)	3.192 [1.721, 6.232]	0.000	0.89 (0.56)	2.434 [0.856, 7.903]	0.112
CRT-7				0.209 (0.045)	1.232 [1.129, 1.348]	0.000	0.132 (0.121)	1.141 [0.9, 1.457]	0.275
Explicit : CRT-7							0.098 (0.139)	1.103 [0.836, 1.451]	0.481
Implicit : CRT-7							0.08 (0.139)	1.083 [0.821, 1.423]	0.568

Note. Each logistic regression model mentioned in this section is summarized in this table with values listed for the variables analyzed in each model. Displayed *p*-values are based on Wald statistics.

To test if the implicit condition was statistically equivalent to the control group, an equivalence test comparing the proportion of correct Linda responses in each group was performed using Fisher's exact *z*-test. The null equivalence hypothesis was that the absolute difference in the proportion of implicit condition participants answering the Linda problem correctly was 15% or more compared to the control group. Fisher's exact *z*-test failed to reject this null equivalence hypothesis, ($z = 1.322$, $p = 0.907$), meaning that the difference in the proportion of correct Linda responses between the implicit and control prompt groups

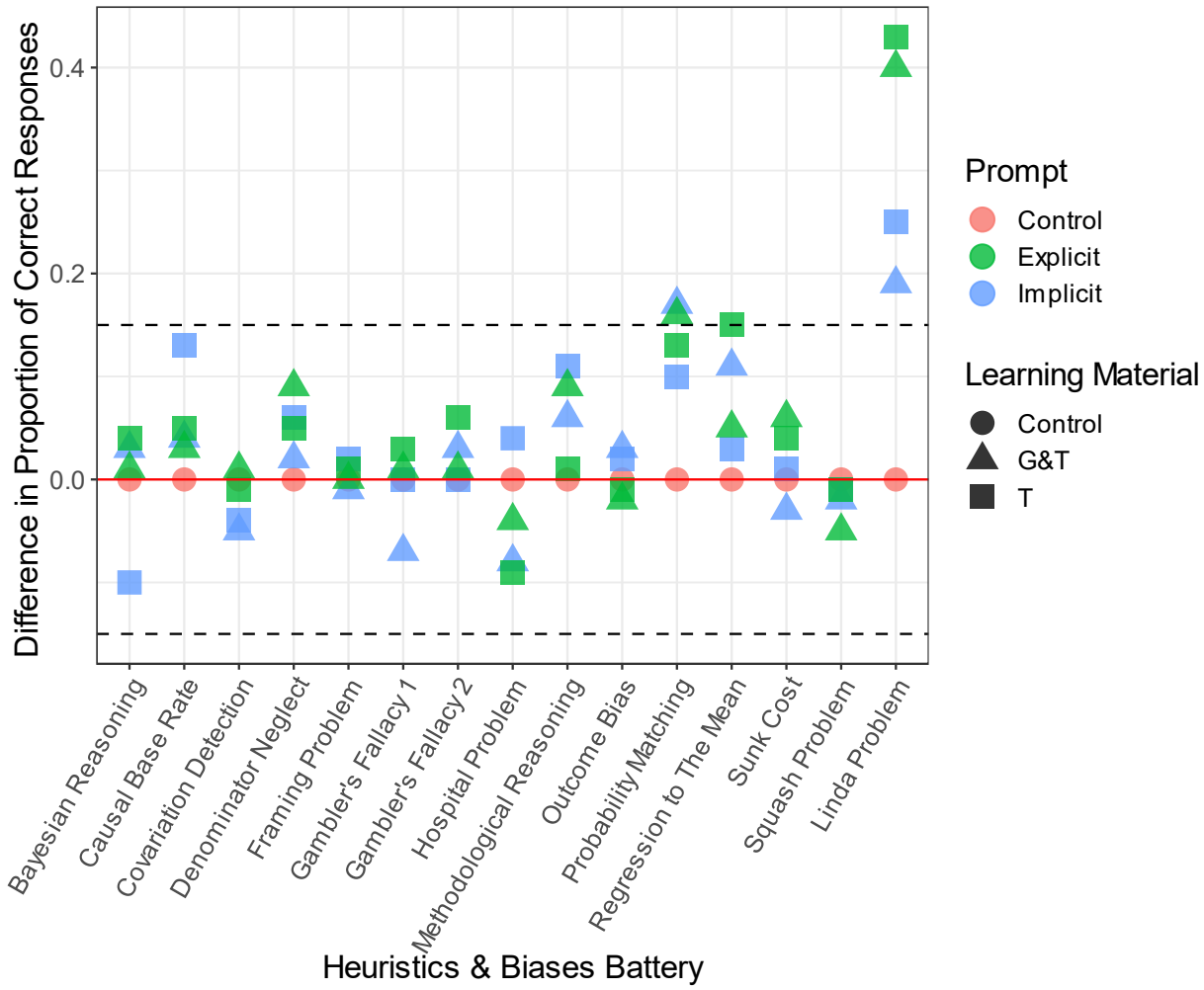
(proportion difference = 0.219, 90% CI [0.133, 0.305]) were not statistically equivalent under the 15% criterion.

Comparing Content and Training Itself

I also conducted an exploratory analysis to infer whether it was specifically the content of the training material or the act of being trained itself that created this training effect. To do this, I graphed the difference in the proportion of correct responses on every question of the heuristics and biases battery between every training group to the control group in Figure 7. The largest increases in performance were found in the Linda problem, particularly for the explicit condition, but less-so for the implicit. Furthermore, most of the proportion of correct responses in each training group for every non-Linda problem falls within the 15% criterion. Notably, trained participants consistently come close to the 15% difference for the probability matching problem.

Figure 7

Simple Comparison of Proportion of Correct Responses on Each Heuristics and Biases Battery Question Between Every Training Group to the Control



Note. Dashed lines represent the 15% difference from control group criterion and the red horizontal line represents the control group for each question. The proportion of correct responses for each question for every non-control Prompt X Learning Material condition was subtracted from the proportion of correct responses in the control group. 'G&T' stands for learning materials with graphs and text. 'T' stands for text only.

Comparing Transfer Effects While Controlling for Reflectivity

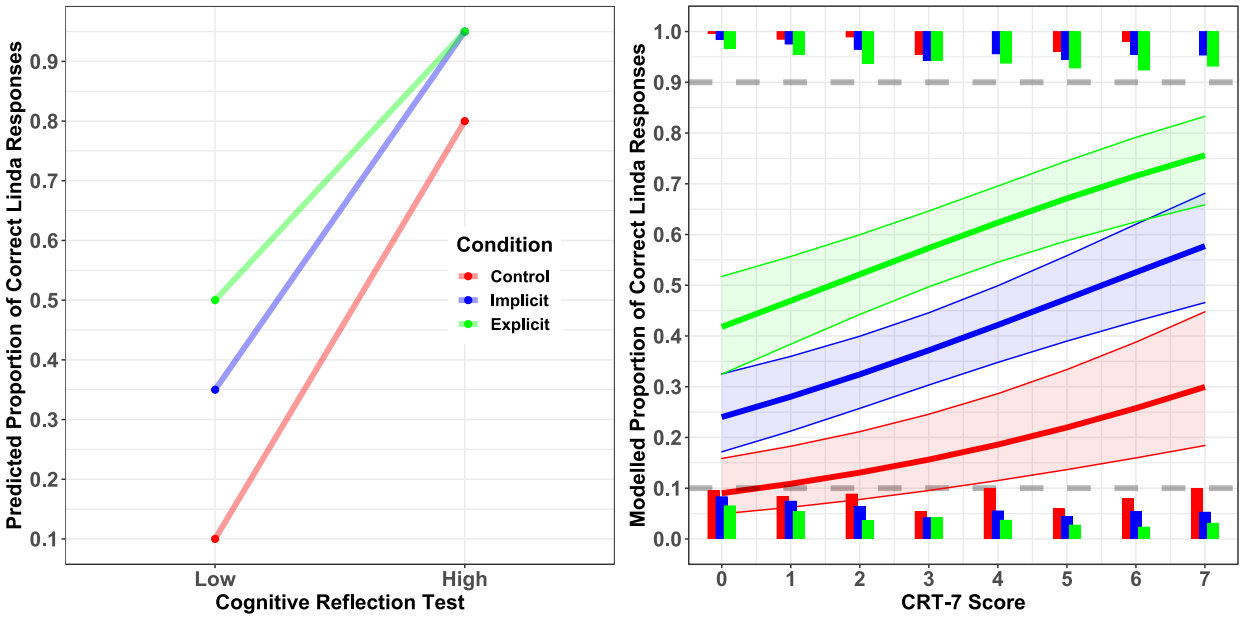
As aforementioned, the CRT has been reported as the strongest predictor of performance on heuristics and biases problems (Toplak et al., 2011). Hence, as shown in Table 5, to infer whether transfer occurs even after controlling for important individual differences, two more

logistic regressions were conducted to compare the prompt conditions while controlling for CRT-7 scores. The first of these two models analyzes the additive effect of CRT-7 scores and the second additionally models an interaction between CRT-7 scores and the prompt condition. As shown in Table 5, using Wald tests, the interaction between CRT-7 and prompt were both statistically insignificant for both the implicit and explicit prompt dummy variables: $p = 0.568$ and $p = 0.481$, respectively. A likelihood-ratio test of the interaction also supporteded this statistically insignificant interaction ($\chi^2 = 1.411$, $df = 2$, $p = 0.494$). Hence, analysis focused on the additive model and is reported here.

Computing the MEM and similar to when only the prompt condition was modelled, when controlling for CRT-7 scores, the odds of answering the Linda problem correctly were significantly greater for each trained group compared to the control group (*OR* = 3.19, 95% CI [1.68, 6.05] for the implicit condition and *OR* = 7.25, 95% CI [3.81, 13.78] for explicit). Performance in the explicit condition was also significantly greater ($p < 0.001$) than in the implicit condition (*OR* = 2.27, 95% CI [1.47, 3.51]), after controlling for CRT-7 scores. Lastly, using Wald tests and as shown in Table 5, CRT-7 was a significant ($p < 0.001$) positive predictor of performance on the Linda problem. Averaged across the prompt conditions, the AME for participants with CRT-7 scores one standard deviation above the mean answered the Linda problem correctly 20.95% more (95% CI [12.41, 29.50]) than CRT-7 scorers one standard deviation below the mean. Figure 8 illustrates the results of this regression model of the additive effects of prompt and CRT-7 side-by-side with the original predictions made for objective 1.

Figure 8

Side-by-Side Comparison of Objective 1 Predictions and Multiple Logistic Regression Model of Additive Effects of Prompt and CRT-7 on Correct Linda Responses



Note. On the left is the prediction for the proportion of correct responses in each prompt condition based on CRT scores and on the right is what is modelled based on the collected data. Vertical bars represent the raw proportion of either correct (top bars) or incorrect (bottom bars) Linda responses for each prompt condition at each CRT-7 score, similar to the style of a logistic histogram (Smart et al., 2004). If bars reach the dashed grey lines at the 0.1 and 0.9 levels of the y-axis, then a 100% of the participants in that condition at that CRT-7 score performed the Linda problem incorrectly or correctly, respectively. Also, the specific values on the y-axis are uninformative in interpreting these vertical bars. Shaded area represents 95% confidence intervals around each regression line.

Objective 2: What Role do Individual Differences Play in Transfer Success?

Confirmatory: Reflectivity and Cognitive Ability

To test my predictions regarding the differential importance of reflectivity—operationalized as CRT-7 scores—and cognitive ability—operationalized as Verbal-Numerical Reasoning (VNR) scores—based on prompt condition, a logistic regression model of accuracy on the Linda problem predicted by a three-way interaction—and its lower-order effects—

between prompt, CRT-7 and VNR was analyzed. As shown in Table 6, based on a Wald test, there was a statistically significant 3-way interaction for the explicit prompt, CRT-7 and VNR ($p < 0.05$). However, performing the same analysis using the sensitivity dataset of 280 participants did not show the same 3-way interaction ($p = 0.150$), as shown in Table 6. Similarly, a more reliable technique for statistical significance using a likelihood ratio test of the 3-way interaction between prompt, CRT-7 and VNR also showed no significant interaction ($\chi^2 = 2.716$, $df = 2$, $p = 0.257$), in the baseline 453 participant dataset. Similarly, there were no statistically significant 2-way interactions throughout analyses of the baseline and sensitivity datasets. Hence, the results in this section are based on a regression model of the additive effects of prompt condition, CRT-7 and VNR scores, as shown in the first set of non-variable columns in Table 6 under the heading 'Prompt, CRT-7 & VNR Additive'.

Table 6

Objective 2 Confirmatory Regression Models

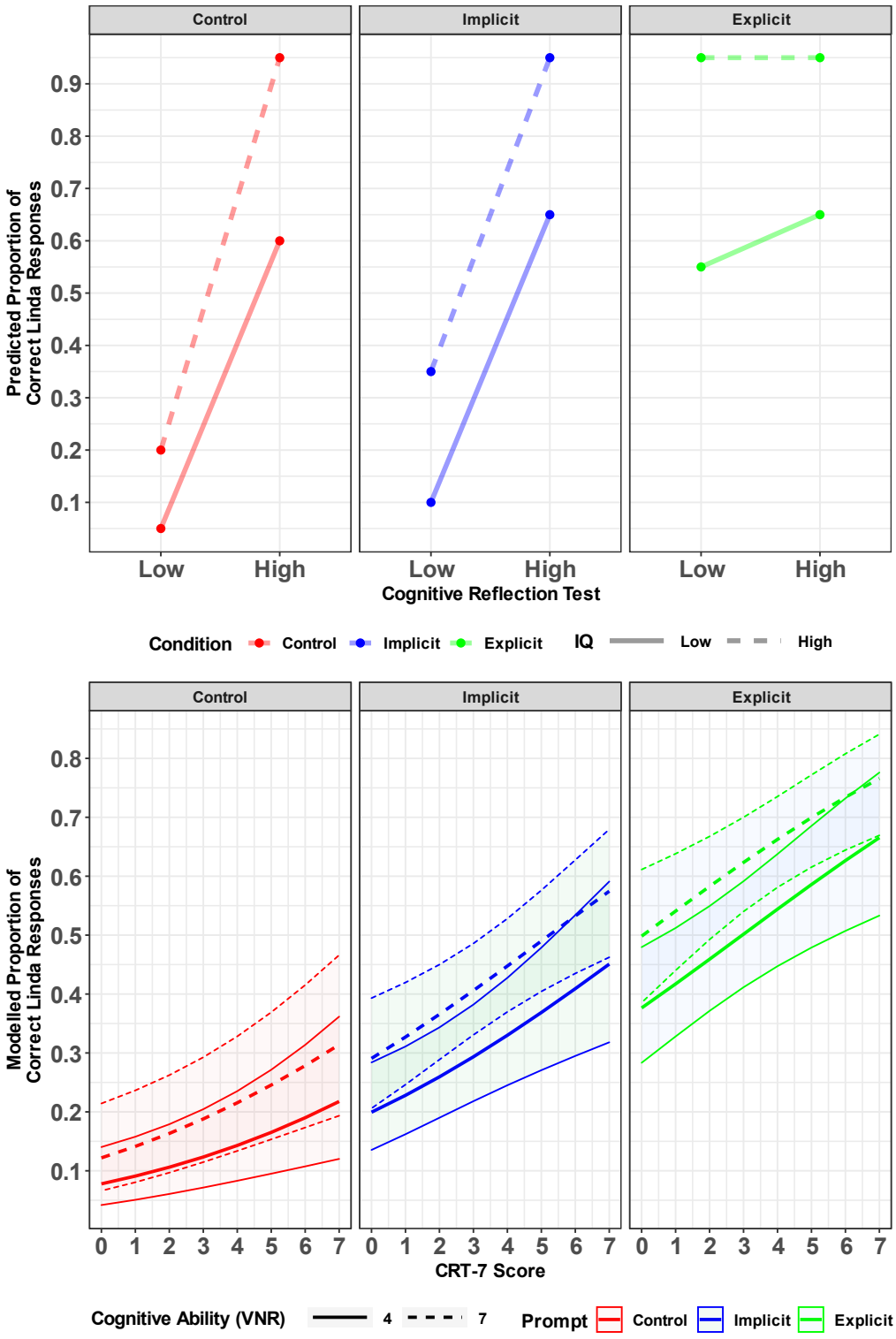
Variable	Prompt, CRT-7 & VNR Additive			3-way Interaction Baseline			3-way Interaction Sensitivity		
	β (SE)	OR 95% CI	p	β (SE)	OR 95% CI	p	β (SE)	OR 95% CI	p
Control-Intercept	-3.136 (0.446)	0.043 [0.018, 0.101]	0.000	-4.006 (1.842)	0.018 [0, 0.433]	0.030	-4.708 (3.582)	0.009 [0, 3.56]	0.189
CRT-7	0.17 (0.047)	1.186 [1.082, 1.302]	0.000	0.867 (0.487)	2.38 [0.986, 6.805]	0.075	1.334 (1.018)	3.796 [0.665, 38.557]	0.190
Explicit	1.967 (0.329)	7.15 [3.839, 14.057]	0.000	2.886 (2.035)	17.921 [0.466, 1510.221]	0.156	3.456 (3.78)	31.68 [0.048, 196992.583]	0.361
Implicit	1.082 (0.329)	2.951 [1.582, 5.787]	0.001	2.178 (2.02)	8.828 [0.236, 724.61]	0.281	2.184 (3.759)	8.878 [0.014, 53376.771]	0.561
VNR	0.166 (0.058)	1.18 [1.056, 1.324]	0.004	0.362 (0.314)	1.436 [0.801, 2.8]	0.250	0.583 (0.59)	1.791 [0.622, 6.795]	0.323
CRT-7 : VNR				-0.133 (0.085)	0.875 [0.729, 1.022]	0.119	-0.234 (0.174)	0.791 [0.531, 1.061]	0.177
Explicit: CRT-7				-1.01 (0.548)	0.364 [0.115, 1.003]	0.065	-1.34 (1.069)	0.262 [0.024, 1.696]	0.210
Explicit : CRT-7:VNR				0.194 (0.097)	1.214 [1.015, 1.485]	0.045	0.262 (0.182)	1.3 [0.949, 1.962]	0.150
Explicit : VNR				-0.223 (0.352)	0.8 [0.385, 1.554]	0.526	-0.382 (0.626)	0.683 [0.171, 2.132]	0.542
Implicit : CRT-7				-0.653 (0.537)	0.52 [0.168, 1.405]	0.224	-0.962 (1.064)	0.382 [0.035, 2.454]	0.366
Implicit : CRT-7 : VNR				0.127 (0.093)	1.136 [0.957, 1.379]	0.169	0.202 (0.18)	1.224 [0.899, 1.84]	0.260
Implicit : VNR				-0.232 (0.346)	0.793 [0.386, 1.52]	0.501	-0.34 (0.618)	0.712 [0.18, 2.184]	0.583

Note. Similar to Table 5, coefficients with their corresponding standard errors and 95% confidence intervals were natural exponentiated to compute the shown odds ratio and its accompanying statistics. Displayed *p*-values are based on Wald statistics.

Similar to objective 1 results, controlling for both CRT-7 and VNR scores through the MEM, the odds of correct Linda responses were significantly greater ($p < 0.001$) for each trained condition compared to the control group ($OR = 2.95$, 95% CI [1.55, 5.63] for the implicit condition and $OR = 7.15$, 95% CI [3.75, 13.64] for explicit). Additionally, controlling for CRT-7 and VNR scores, explicit prompt participants performed significantly better ($p < 0.001$) than the implicit ($OR = 2.42$, 95% CI [1.56, 3.77]). Averaged across the prompt conditions and VNR scores to compute the AME, there was a difference of 16.74% (95% CI [7.76, 25.72]) in the proportion of correct Linda responses between participants with CRT-7 scores one standard deviation above and below the mean CRT-7 score. Averaged across the prompt conditions and CRT-7 scores, there was a difference of 13.45% (95% CI [4.32, 22.58]) in the proportion of correct Linda responses between participants with VNR scores one standard deviation above and below the mean VNR score. A diagram of these results, its corresponding regression model and my predictions are shown in Figure 9. In summary, the training effects in the implicit and explicit groups reported in objective 1 were still significant even after controlling for both CRT-7 and VNR scores. Also, even after controlling for other factors, an increase in either CRT-7 or VNR scores both predicted better performance on the Linda problem. However, the predicted interactions between individual differences and training were not consistently statistically significant in my analysis.

Figure 9

Comparison of Objective 2 Confirmatory Hypotheses and Multiple Logistic Regression Model of Additive Effects of Prompt, CRT-7 and VNR



Note. To differentiate between high and low cognitive ability, shown are the model estimates roughly based on the 75th and 25th percentile scores (see Appendix D for details) of the VNR: 7 out of 13 and 4 out of 13, respectively. Shaded around each regression line is the 95% confidence interval.

Exploratory: Recognition Memory

To test my predictions regarding the differential importance of recognition memory—operationalized as *d*-prime scores from a verbal paired associates recognition task—I ran a logistic regression model of Linda problem accuracy predicted by a two-way interaction between prompt and *d*-prime combined with the additive effects of VNR and CRT-7. As shown in Table 7 under 'Prompt & dprime Interaction' and computed using Wald tests, there was no significant 2-way interaction between *d*-prime and the implicit prompt ($p = 0.239$) nor *d*-prime and the explicit prompt ($p = 0.436$). Likelihood ratio tests of the interaction were also statistically insignificant ($\chi^2 = 1.433$, $df = 2$, $p = 0.488$). Hence, the results in this section are based on the additive effects of prompt condition, CRT-7, VNR and *d*-prime scores, as shown in Table 7 under the heading 'Prompt, CRT-7, VNR and dprime Additive'.

Table 7

Objective 2 Exploratory Regression Models

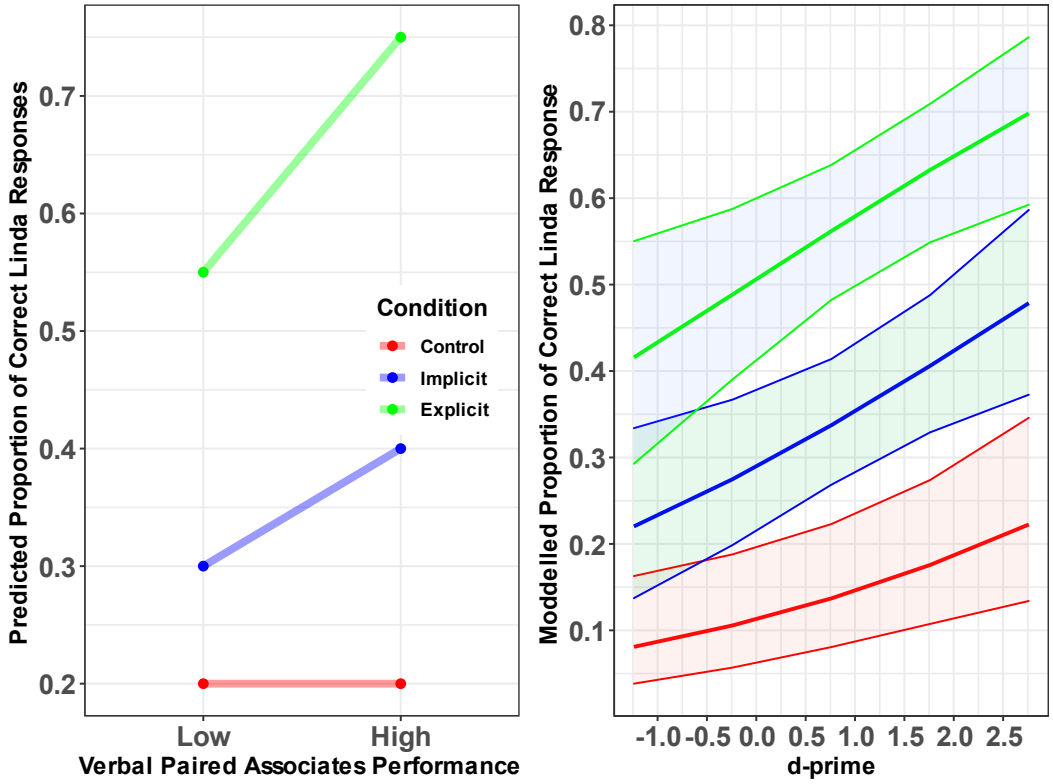
Variable	Prompt, CRT-7, VNR & dprime Additive			Prompt & dprime Interaction		
	β (SE)	OR 95% CI	*p*	β (SE)	OR 95% CI	*p*
Control-Intercept	-3.437 (0.466)	0.032 [0.012, 0.078]	0.000	-3.056 (0.574)	0.047 [0.014, 0.138]	0.000
CRT-7	0.148 (0.048)	1.159 [1.055, 1.275]	0.002	0.148 (0.048)	1.159 [1.055, 1.275]	0.002
Explicit	2.09 (0.337)	8.084 [4.28, 16.142]	0.000	1.723 (0.522)	5.603 [2.11, 16.616]	0.001
Implicit	1.166 (0.334)	3.208 [1.705, 6.349]	0.000	0.646 (0.538)	1.907 [0.691, 5.803]	0.231
VNR	0.155 (0.058)	1.168 [1.043, 1.311]	0.008	0.156 (0.058)	1.168 [1.043, 1.312]	0.008
dprime	0.295 (0.1)	1.343 [1.107, 1.637]	0.003	0.043 (0.263)	1.044 [0.612, 1.742]	0.871
Explicit : dprime				0.234 (0.301)	1.264 [0.704, 2.314]	0.436
Implicit : dprime				0.36 (0.306)	1.434 [0.791, 2.647]	0.239

Note. Done in the same manner as Table 6.

Based on the additive model and averaging across prompt conditions, CRT-7 and VNR scores to compute the AME, there was a difference of 13.13% (95% CI [4.52, 21.75]) in the proportion of correct Linda responses between participants with *d*-prime scores one standard deviation above and below the mean *d*-prime score. A diagram of these results, its corresponding regression model, and my predictions are shown in Figure 10. In summary, controlling for training, CRT-7 and VNR scores, higher *d*-prime scores predicted better performance performance on the Linda problem. However, the predicted interaction between training and *d*-prime scores was not statistically significant.

Figure 10

Comparison of Exploratory Objective 2 Predictions and Multiple Logistic Regression Model of Additive Effects of Prompt, CRT-7, VNR and d-prime



Note. Regression lines are made controlling for CRT-7 and VNR scores. Done in the same manner as Figure 9.

Objective 3: Does the Kind of Learning Material Matter?

Statistical Equivalence Between Kinds of Learning Materials

To test if learning materials with graphs and text were statistically equivalent to materials with only text, I ran an equivalence test using Fisher's exact *z*-test. Similar to what was done in objective 1, equivalence was defined as when the proportions of correct responses are within 15% of each other. Combining the implicit and explicit prompt conditions together, Fisher's exact *z*-test rejected this null equivalence hypothesis, ($z = 1.912$, $p < 0.05$), the difference in proportion of correct Linda responses between the learning material conditions with graphs and

text versus only text was -4.9% (90% CI [-13.6, 3.70]). Furthermore, using null hypothesis significance testing, Fisher's exact *z*-test failed to reject the null significance hypothesis (z = -0.940, p = 0.347). In other words, the learning materials were (1) statistically equivalent and also (2) not statistically different from each other in the baseline analysis.

However, using the sensitivity dataset of 280 participants, the null equivalence hypothesis failed to be rejected, (z = 0.151, p = 0.440), with a difference of -14% (90% CI [-0.245, -0.036]), favoring text only. Additionally, the null significance hypothesis was rejected, (z = -2.204, p = 0.0275) using Fisher's exact *z*-test, favoring text only. In other words, with the sensitivity dataset, the learning materials (1) were not statistically equivalent and (2) were statistically different from each other.

In summary, the statistical conclusions of whether the kinds of learning materials are (1) statistically equivalent and (2) statistically different from each other are opposites between the baseline and sensitivity datasets. To probe the issue further, I analyzed the importance of the prompt condition and individual differences using logistic regression and null hypothesis significance testing.

Statistical Difference Between Kinds of Learning Materials While Accounting for Individual Differences

To analyze whether the learning materials were statistically different from each other while accounting for individual differences, I exploratorily created a series of multiple logistic regressions building from using only prompt and learning material as predictors, to incorporating CRT-7, VNR and *d*-prime scores. Ultimately, guided by the aforementioned principles of parsimony and fit statistics based on AIC, BIC and likelihood-ratio statistics, I settled on the model shown in Table 8 and also compared my baseline results with the sensitivity dataset.

Table 8

Objective 3 Logistic Regression Table

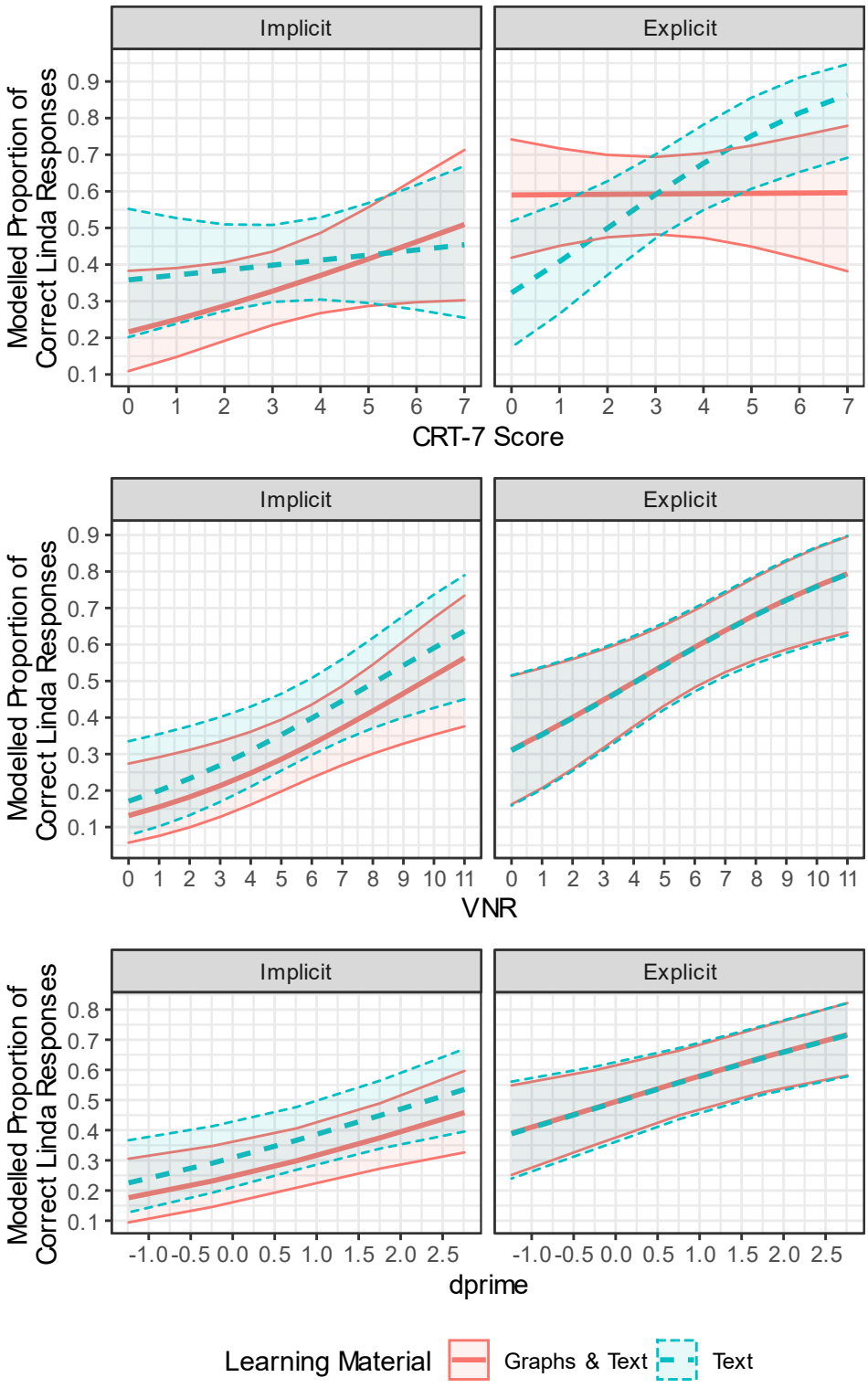
Variable	Baseline			Sensitivity		
	β (SE)	OR 95% CI	*p*	β (SE)	OR 95% CI	*p*
Implicit Graphs & Text (Intercept)	-2.853 (0.555)	0.058 [0.018, 0.164]	0.000	-3.181 (0.727)	0.042 [0.009, 0.16]	0.000
CRT-7	0.189 (0.103)	1.208 [0.991, 1.49]	0.066	0.25 (0.135)	1.284 [0.994, 1.698]	0.064
Explicit	1.652 (0.54)	5.217 [1.856, 15.582]	0.002	1.725 (0.698)	5.614 [1.487, 23.467]	0.013
Explicit : CRT-7	-0.186 (0.137)	0.83 [0.632, 1.085]	0.176	-0.233 (0.172)	0.792 [0.559, 1.105]	0.176
Explicit : Text	-1.81 (0.79)	0.164 [0.034, 0.76]	0.022	-1.563 (1.017)	0.209 [0.028, 1.515]	0.124
Explicit : Text : CRT-7	0.498 (0.206)	1.646 [1.103, 2.48]	0.016	0.593 (0.267)	1.809 [1.083, 3.105]	0.027
VNR	0.195 (0.065)	1.215 [1.072, 1.383]	0.003	0.196 (0.081)	1.216 [1.04, 1.433]	0.016
dprime	0.344 (0.111)	1.411 [1.138, 1.763]	0.002	0.313 (0.14)	1.367 [1.045, 1.811]	0.025
Text	0.705 (0.575)	2.023 [0.659, 6.377]	0.221	1.358 (0.736)	3.887 [0.941, 17.292]	0.065
Text : CRT-7	-0.132 (0.144)	0.876 [0.658, 1.161]	0.358	-0.262 (0.181)	0.769 [0.535, 1.091]	0.146

Note. Done in the same manner as Table 7.

As shown in Table 8, I found a statistically significant 3-way interaction between prompt, learning material and CRT-7 scores in both the baseline and sensitivity datasets ($p = 0.027$), using Wald tests. This statistical significance was consistent using the more reliable likelihood ratio tests in both the baseline ($\chi^2 = 5.97$, df = 1, $p < 0.05$) and sensitivity ($\chi^2 = 5.14$, df = 1, $p < 0.05$) datasets. Lastly, the additive effects of VNR and *d*-prime were each statistically significant across both datasets using both Wald tests and likelihood ratio tests ($p < 0.05$). The baseline regression model is depicted in Figure 11 before discussing specific effect sizes. The sensitivity regression model is graphed in Figure 15 in Appendix D.

Figure 11

Visualization of Objective 3 Baseline Regression Model



Note. Each row of graphs represents a different set of coefficients being varied at the same time. For example, the first row varies CRT-7 scores, prompt and learning conditions simultaneously, while controlling for *d*-prime and VNR scores. Similarly, the second row varies VNR scores, prompt and learning conditions simultaneously, while controlling for *d*-prime and CRT-7 scores. Note that although the 2nd and 3rd rows may be misleading due to using a set CRT-7 score given that CRT-7 is part of a 3-way interaction with prompt and learning material, the plots are still roughly consistent with the average marginal effects detailed in the text. Shaded areas represent 95% confidence intervals.

Beginning with the main effects and similar to what was found in objective 2, as either VNR or *d*-prime increases, the probability of answering the Linda problem increases. Averaging across prompt conditions, CRT-7 and *d*-prime scores to compute the AME, there was a difference of 16.48% (95% CI [5.83, 27.13]) in proportion of correct Linda responses between participants with VNR scores one standard deviation above and below the mean VNR score. Averaging across prompt conditions, CRT-7 and VNR scores, there was a difference of 16.34% (95% CI [6.08, 26.6]) between participants with *d*-prime scores one standard deviation above and below the mean *d*-prime score.

To summarise the three way interaction, the differences in the proportion of correct Linda responses at each learning material and prompt level were compared between CRT-7 scorers one standard deviation above and below the mean CRT-7 score, while averaging across VNR and *d*-prime scores. The results of these comparisons are shown in Table 9. Importantly, the increase in proportion of correct responses between CRT-7 scorers one standard deviation above and below the mean in the explicit graphs and text is very small at 0.39% (95% [-19.52, 20.30]), but the increase in the explicit text condition is much larger at 36.73% (95% CI [16.11, 57.36]). However, these differences within the explicit condition are not as dramatic within the implicit condition.

Table 9

Effects of CRT-7 at Each Level of Learning Material and Prompt

Prompt	Learning Material	CRT-7 AME	Difference in Proportions (%)	SE	95% CI	
					LL	UL
Implicit	Graphs & Text	(x + sd) - (x - sd)	18.352	9.858	-0.969	37.674
Implicit	Text	(x + sd) - (x - sd)	6.007	10.954	-15.462	27.475
Explicit	Graphs & Text	(x + sd) - (x - sd)	0.392	10.157	-19.515	20.300
Explicit	Text	(x + sd) - (x - sd)	36.734	10.522	16.111	57.356

Note. 'x' represents the mean CRT-7 score in the entire baseline dataset that was analyzed, meaning without control group participants. CRT-7 AME stands for CRT-7 average marginal effect, meaning the effect of CRT-7 averaged across all other non-specified variables. In this case, that means averaged across VNR and *d*-prime scores for each prompt and learning material level. SE stands for standard error of the difference, and LL and UL correspond to the lower and upper limits of the 95% CI, respectively.

Discussion

Objective 1: Dual-Process Compliant Transfer Occurs Even After Accounting for Individual Differences

The first and primary objective of this study was to revisit prior transfer studies on the inhibition of heuristics (e.g. Agnoli, 1991; Agnoli & Krantz, 1989; Fisk & Pidgeon, 1997) under the lens of dual process theory: to investigate whether transfer under the terms of dual process theory occurs or not. Firstly, I successfully replicated the finding that approximately 80% of untrained participants make a conjunction error on the Linda problem (Agnoli & Krantz, 1989; Fiedler, 1988; Hertwig & Chase, 1998; Tversky & Kahneman, 1983), as shown in the cell for control group participants in Table 4. Secondly, I successfully replicated the large training effects reported by Agnoli and Krantz (1989): when participants are explicitly prompted or

strongly suggested to apply their learning, there was an improvement of about 40% compared to the control, as shown in Table 4.[11] Lastly, even when the relationship between the learning material and the Linda problem is obscured by (1) embedding the Linda problem within a battery of other unrelated heuristics and biases problems; (2) temporally separating the training from presentation of the Linda with a substantial number of intervening tasks; and (3) not explicitly prompting participants to apply their learning—there is still a statistically significant positive effect of training, as shown in Table 5. Moreover, using equivalence tests, I found that the control and implicit prompt groups were not statistically equivalent under a 15% criterion. As shown in Table 4, conservatively, there was about a 20% improvement in the implicit group compared to the no-training control group.

Further strengthening the case for transfer success, I found that training produced strong and positive effects specifically to the Linda problem and not to any other heuristics and biases battery problem, as shown in Figure 7. This suggests that transfer effects were a result of the content of the training specifically, and not some general training effects—i.e., the act of being trained. If there were also strong positive effects on many other heuristics and biases problems unrelated to the training—i.e., meaning non-Linda problems—then it may have been the case that the act of being trained alerted people to be particularly careful on all other questions. In this hypothetical scenario, higher performance across the heuristics and biases battery could be attributed to reducing biased responses from system 1 process, meaning reducing cognitive miserliness, (Stanovich et al., 2016; Toplak et al., 2011), and not the content of the training itself. One may argue that the training content may also be transferable to other heuristics and biases

[11] It should be noted that Agnoli and Krantz (1989) reported the mean proportion of errors of a set of conjunction-error questions and not the individual proportion of errors on the Linda problem. Nevertheless, the large difference between explicitly trained and control group participants is still consistent.

problems—such as the probability matching and regression to the mean problems, as shown in Figure 7. However, given that I had no theory-based prior belief in the broad applicability of the training materials used here and the generally unsubstantial differences in correct responses compared to transfer success on the Linda problem, I consider such a possibility unlikely.

Moreover, even after controlling for established individual differences thought to importantly predict performance on heuristics and biases tasks (Stanovich et al., 2016; Toplak et al., 2011), I found that participants in the implicit condition still performed significantly better on the Linda problem than those in the control group, as shown in Tables 5 and 6, then illustrated in Figures 8 and 9. Importantly, I did not find a statistically significant interaction between prompts and any individual differences, suggesting that training benefited all participants comparably. Similar to when individual differences were not accounted for, participants who were told to apply their learning performed significantly better than when the relationship between the learning material and the problem was masked.

Taking these findings together, there is strong reason to believe that transfer success under dual process theory on the Linda problem does occur, and that this training benefits everyone, regardless of individual differences in reflectivity—the strongest known predictor of performance on questions like the Linda problem (Toplak et al., 2011). On the basis of both null hypothesis significance and equivalence testing, I consistently found that even when the purpose of the training material was masked, transfer success was identified. In other words, the absence of statistically non-significant interactions between prompt and individual differences in the results for either objectives 1 and 2, imply that training benefitted everyone, regardless of the individual differences measured here—reflectivity, cognitive ability and recognition memory. A discussion of (1) the failure to find a statistically significant interaction between CRT-7 and

prompt condition; and (2) a confirmation of the positive effects of better reflectivity follows in the second objective. Generalizing these results, there is now an even stronger reason to believe that education truly can inhibit our heuristics, even when umprompted (cf. Yarkoni, 2022). Although one may critique this conclusion on the grounds that the experiment was conducted online—while citing the complexities of online research—it is also the case that this experiment replicated prior key findings conducted in-person (Agnoli & Krantz, 1989; Fisk & Pidgeon, 1997). Furthermore, these claims are bolstered by accounting for important individual differences identified by tripartite theory (Stanovich, 2018; Stanovich et al., 2016)—a developed form of dual-process theory. In other words, critiquing these findings would also require a critique of prior in-person studies, which may also entail an examination of in-person studies in general.

Objective 2: Individual Differences in Reflectivity, Cognitive Ability and Recognition Memory Predict Higher Performance Regardless of Training Condition

As shown in Figure 9 and Table 6, I failed to falsify the null hypothesis for the differential effects of reflectivity and cognitive ability based on prompt predicted by the tripartite model, specifically Figure 3.3 of Stanovich et al. (2016). In retrospect, one explanation for this failure in identifying a three way interaction may be due to a combination of insufficient statistical power, the high correlations between the CRT and cognitive ability tests like the VNR (Toplak et al., 2011, 2013), and the strong conceptual interconnectedness between individual differences in thinking dispositions and fluid intelligence. Despite this, I was able to conceptually replicate Toplak et al. (2011) by finding statistically significant and strong positive effects for both reflectivity and cognitive ability on performance on the Linda problem, as shown in Figure 9 and Table 6.

However, given the high correlations between reflectivity and cognitive ability, as shown by Toplak et al. (2011, 2013) and replicated here (Figures 16 and 17 in Appendix D), it is interesting to note that a substantial proportion of the highest scorers of the CRT-7 also answered the Linda problem incorrectly, as shown in the raw data visualization in Figure 8. Furthermore, the lack of any interaction is puzzling because it implies that increases in reflectivity and cognitive ability should predict comparable increases in correct responses on the Linda problem in both trained and untrained participants. Furthermore, this lack of interaction was found even in the sensitivity analysis where all trained participants must have answered all learning material questions correctly. Hence, similary to Szaszi et al. (2017), I argue that the CRT-7 may measure more than just dispositional reflectivity. However, given that cognitive ability was already controlled for using the VNR task, I argue that it may be the case that there is some other important missing construct to this problem that is being partly measured, but not understood.

In a similar manner, my exploratory incorporation of measures of recognition memory through the VPA task found a significant positive effect of recognition performance, even after controlling for prompt condition, CRT-7 and VNR scores, as shown in Table 7 and Figure 10. However, I also failed to falsify the null two way interaction, finding little support for the two way interaction that I predicted on the basis that memory measures should only be strongly positively associated with better Linda performance in the trained groups and not the control. In other words, my prediction that individual differences in recognition memory should have a stronger effect in the trained groups and no effect in the no-training control group because those with better recognition memory should be better able to apply their training, was not supported. I had no prior reason to predict that better memory should entail better performance regardless of training.

One potential reason why my predictions regarding an interaction between recognition memory and prompt condition were not supported may be due to the validity of using VPA in this scenario. Here, VPA tests people's ability to remember the exact match between learned and tested word pairs. However, transfer of learning to the Linda problem is not so much an exact match of the learning material and the problem, but rather a match in the underlying cognitive content. Therefore, using VPA as a measure of individual differences in transfer may be partly invalid. Nevertheless, it may be valid in the sense that the VPA is a measure of people's ability to learn and remember information and that people's ability to perform these tasks—learning and remembering—is still relevant to transfer of learning. Future research should explore the VPA further in the context of learning and transfer ability, while also exploring other measures of memory to further clarify the situation. To further compound this perplexity, it's interesting to note that Gigerenzer and Gaissmaier (2011) argued that the use of heuristics in biased judgments is adaptive and theorized that recognition memory is an important component of making adaptive judgments. However, here I found that having more accurate recognition memory predicts a higher chance of answering the Linda problem correctly—or 'unbiasedly'—even without training.

In summary, I found that reflectivity, cognitive ability and recognition memory, operationalized by CRT-7, VNR and *d*-prime scores, respectively, are all unique positive predictors of performance on the Linda problem, even while controlling for each other and across all prompt conditions. This also represents a successful replication of the finding that CRT and cognitive ability are significantly correlated with performance on the heuristics and biases task of Toplak et al. (2011). However, none of my predictions regarding interactions between individual differences and the prompt conditions were strongly supported by my analyses.

Objective 3: It is Unclear if Transfer Success Differs Between Kinds of Learning Materials and Learning Materials may Accentuate Individual Differences Under Certain Contexts

Without accounting for prompt or individual differences, it is unclear whether training effects of graphs and text compared to text-only are (1) statistically equivalent and (2) statistically different. The baseline dataset suggests that the two are statistically equivalent and not statistically different, while the sensitivity dataset suggests the opposite. However, if there is a true difference between the materials, the analyses suggest that purely textual materials are better for transfer success in this specific scenario. To probe the issue further I used multiple logistic regression to investigate the potential role of individual differences and prompts.

As shown in Table 8 and Figure 11, I found that those with better recognition memory and cognitive abilities—as measured by *d*-prime and VNR, respectively—performed better on the Linda problem. There were no significant differences between learning materials based on either recognition memory or cognitive ability. However, I also found a three way interaction between prompt, learning material and CRT-7 scores. Averaged across *d*-prime and VNR scores, as shown in Table 9 and Figure 11, in the explicit prompt condition, increases in CRT-7 score predict very small increases in Linda response accuracy when learning materials had graphs and text, but very large increases with text-only. The situation reverses with learning material in the implicit prompt condition, but the differences based on an increase of CRT-7 are much smaller. In summary, although higher cognitive ability and recognition memory predict better performance regardless of prompt or learning materials—similar to objective 2—when participants are explicitly prompted, then learning materials with graphs and text practically nullified the importance of reflectivity, whereas learning materials with text-only accentuated it.

Another interpretation of this three way interaction is that in situations where people will be explicitly prompted to apply their learning, then it is best to administer textual learning materials to those high in reflectivity and graphs and text to those lower in it. However, although the three-way interaction replicates in the sensitivity dataset and likelihood ratio tests, I question whether the strength of the experimental design and data is sufficient to warrant such a prescription. Firstly, as shown in Table 8, the *p*-value of the three way interaction is not highly significant, at around $p = 0.02$—averaged across baseline and sensitivity analyses. Hence given the many analyses already conducted in this thesis, it seems unlikely to survive a *p*-value correction. Secondly, there was no prior reason to predict that such an interaction would occur, so making this specific prescription would not be a theory-guided decision. Thirdly—even without statistical concerns—conceptually-speaking, making generalized prescriptions based on the transfer of one set of materials to one specific question is unwarranted (Yarkoni, 2022).

A more conservative interpretation of the three way interaction is that learning materials may sometimes accentuate individual differences under certain circumstances. Although the results of objectives 1 and 2 support the prediction made using tripartite model that better reflectivity will predict higher performance on the Linda problem and replicate the findings reported by Toplak et al. (2011), training with graphs and text appears to have nullified the importance of reflectivity when participants are told to apply their learning. In contrast, using purely textual materials sharply increased the importance of reflectivity. Importantly, however, this accentuation of reflectivity was not predicted when participants were not told explicitly told to apply their learning. Although prior literature has made the argument that optimal learning materials may depend on aptitude (e.g. Pashler et al., 2008), to my knowledge, it has not been previously raised that this importance could change based on whether spontaneous transfer is

required or not. Hence, rather than prescribe optimal learning materials based on individual differences, a stronger interpretation is that it is important to account for spontaneity of transfer and cues to apply learning in future educational studies.

If it is the case that purely textual materials truly are better than a combination of graphs and text, the educational literature on abstract versus concrete learning materials offers one potential interpretation. Contrary to the intuition that highly concrete and imageable learning stimuli are desirable in education because of their memorable and understandable nature (e.g. J. M. Clark & Paivio, 1991), Kaminski et al. (2008, 2013) argued that such learning materials may distract learners and also decrease transferability by making training overly attached to the specific learning materials used (c.f. Bock et al., 2011; Jones, 2009a). Following this argument, it may be the case that even if the graphs were more memorable and understandable to the participants, the visuals also limited the transferability of their learning to the Linda problem.

In summary, although it is unclear how graphical and textual learning materials compare with purely textual learning materials at the group level, an analysis of prompt condition and individual differences suggests a nuanced relationship. Specifically, different learning materials may differentially interact with individual differences and that this interaction may differ based on whether participants are explicitly prompted to apply their learning or not. Furthermore, when participants were told to apply their learning, the use of learning materials with graphs and text practically nullified the importance of reflectivity—the strongest established predictor of performance on heuristics and biases tasks (Toplak et al., 2011). Hence, future studies should address spontaneity of transfer of learning because it is an ecologically important factor that enrichens our understanding of training and individual differences.

General Reflection

Understanding Transfer Success in the Implicit Condition Based on Dual Process Theory

As aforementioned in the discussion for objective 1, there were about 20% fewer conjunction errors in trained participants who received no prompt compared to control group participants, which has been defined here as 'transfer success', under dual process theory. Furthermore, although transfer success was still statistically different from control after accounting for important individual differences, none of the predicted interactions between individual differences and training were supported.

From the perspective of tripartite theory, one interpretation is that transfer success occurred in participants who incorporated the training material well enough, such that they were able to detect a conflict between their intuitive biased responses and what they learned, then were able to correctly transfer their learning and answer the Linda problem. However, this interpretation is complicated by the finding that individual differences in recognition memory predicted higher performance even in the control group where no training was provided. Given that memory ability generally predicts better performance regardless of actual knowledge of the training, this implies that the stated interpretation—that those who showed transfer success did so because they learned the material better than others—is not cleanly supported by the analysis of the data. Furthermore, if it was conflict detection and override due to the learning material that caused an improvement in performance, then there should have been an interaction between individual differences in reflectivity and cognitive ability with learning prompt condition, but no strong interactions were consistently identified. Also, even if statistical significance of the interaction is put-aside, the depiction of the additive model in Figure 9 suggests that there is still a substantial proportion of errors even at high levels of both reflectivity and cognitive ability.

Although one may counterargue that Figure 9 does not show the highest levels of cognitive ability, it seems unreasonable that such high ability is necessary to guarantee correct responses on the Linda problem when participants are explicitly prompted to apply their learning. Additionally, it would not cleanly explain why some participants—trained and untrained—at the lowest levels of cognitive ability and reflectivity still answered the problem correctly.

Nevertheless, the exploratory analysis in Figure 7 suggests that it was the content of the training material itself that improved performance, because training selectively and strongly improved performance on the Linda problem, but not more generally on the other problems in the heuristics and biases battery. Therefore, it may be the case that the difficulties in finding clear support for an interpretation of transfer success in the implicit condition using tripartite theory may be due to issues or limitations of the measures used here, as aforementioned. However, the findings in the explicit condition further complicate any simple interpretation under dual process theory and its extension, tripartite theory.

Results for Prompt versus no-Prompt Partially Agrees With Dual Process Theory

The benefits of receiving a prompt over not receiving one agrees with dual process theory: people may realize that their rapid and intuitive judgments are incorrect upon further in-depth, slow and rigorous thought (Kahneman, 2011). However, to my knowledge, there is currently no simple explanation for why approximately 40% of trained participants—which again replicates a result of Agnoli & Krantz (1989)—who were explicitly told that what they learned would be applied to a later problem and also told to apply their learning immediately before seeing the Linda problem, still made conjunction errors. By telling people to apply their learning, presumably, the problem of conflict detection and even how well instantiated the mindware is should be nullified because there should be no need for effortful consideration of

what the correct answer is against one's intuitions. Furthermore, as shown in Table 15 in Appendix D, a similarly large proportion of errors remains even in the sensitivity analysis where trained participants—of which there are many, as shown in Table 1—must have answered all learning material questions correctly. Therefore, this implies that the large proportion of errors that remain in the explicit condition did not occur simply because participants did not understand the training material.

At this point, one may consider whether it may be the training material itself that is problematic and leading to this large proportion of errors. It could be the case that the content of the training material itself maps poorly onto the Linda problem or that it may mislead people into committing conjunction errors. However, referring to my adaptation of the learning material in Appendix C of this thesis and the original materials in Appendix A of Agnoli and Krantz (1989), I argue that there is almost an overinsistent focus on the likelihood of conjunctions versus their constituents in the training materials. Every example in the training material consistently emphasizes that subsets should be less likely to occur than the larger subset they belong to. Furthermore, my adaptation attempted to improve the design language of the original materials by consistently and clearly delineating what is and is not a subset in the graphs and text learning materials. Therefore, I argue that it is unlikely that issues with the learning material are the main cause of the large proportion of errors. Then what about issues specific to the Linda problem?

Over the years, one of the main criticisms of the Linda problem itself is the interpretational ambiguity in what is meant by the 'and' in 'bank teller and feminist' (Hertwig et al., 2008; Mellers et al., 2001). For example, if participants interpreted 'and' as in a logical union to mean that Linda could be a bank teller and/or a feminist, then ranking 'bank teller and feminist' as more likely than its constituents would no longer be a conjunction errors because the

probability that someone is either A and or B is greater than the probability of A or B by itself (Mellers et al., 2001). However, given that explicitly prompted participants in this experiment were taught about conjunctions consistently using the intersectional non-union meaning of 'and', it seems unlikely that participants would interpret it as a union. If it is not about the Linda problem itself or the nature of the learning materials, then could it be that participants disagree with the learning material?

Another argument against interpreting the Linda problem as evidence for people's biased judgments is that it is not universally agreed—it is not a 'Panglossian' truth—that the rules of frequentist probability apply in the case of single frequency events, but rather, such a belief is normative (Gigerenzer, 1991; Stanovich, 2009; c.f. Tversky & Kahneman, 1996). Hence, it may be the case that the large proportion of errors in the explicit condition is due to participants subscribing to a different view of probability. However, I argue that this is unlikely given that the errors remain in the sensitivity analysis where participants answered the simpler conjunction problems in a way that conforms with the normative view. That being said, a future analysis of people's post-experimental justifications for their responses on the Linda problem should provide stronger clarifications towards the nature of their errors.

Overall, this explanatory gap in why there are many errors even in trained participants who are explicitly prompted to apply their learning is consistent with recent commentary by Regenwetter et al. (2022) on the Nobel prize winning work by Tversky & Kahneman (1992): many participants behave in a way that directly contradicts what cumulative prospect theory predicts. There has been general concern that the empirical base related to our theories contains a substantial proportion of participants whose behavior is inconsistent with the theory itself even when the result is statistically significant (Grice et al., 2020). Thus, given that only some of the

participants in both the implicit or explicit condition answered the Linda problem correctly, I speculate that there is some important individual difference that is not being accounted for in this experiment. What is the key individual difference that strongly predicts whether someone will successfully transfer their learning in the implicit group? What about successful transfer in the explicit group?

Spontaneity of Transfer is Underappreciated

Although the review of transfer research by Barnett & Ceci (2002) is very highly cited, their repeated highlighting of the problem of spontaneity in the recognition that a presented problem is related to prior learning—to my knowledge—currently underappreciated in the transfer literature (e.g. Dahlin et al., 2008; Nichols et al., 2021; Owen et al., 2010; Simons et al., 2016; von Bastian et al., 2022), but see the article by Jones (2009b) for a notable exception. While this experiment tackled the problem of spontaneity of transfer as its primary objective, and found spontaneous transfer, I also found that transfer was much improved when participants were explicitly instructed to apply their learning, which aligns with the review of older literature by Barnett & Ceci (2002). In light of the replication crisis and its related siblings (e.g. Chmielewski & Kucker, 2020; Kennedy et al., 2020; Yarkoni, 2022), I argue that this finding is particularly concerning because it implies that transfer studies that do not explicitly account for spontaneity may be either overinflating or even underestimating the benefits of training. (Underestimations may occur when the real world context reliably cues a person's transfer of learning—such as mathematics at the cash register—while the experiment masks the learning's relevance, as done here.) As said before, in the real world—outside of the classroom and laboratory—there rarely are many explicit indicators that and how what we learn in school can be applied to a new problem. Further examples of this underappreciation can be found in: the

educational literature (e.g. Fyfe et al., 2015; Goldstone & Son, 2005; Kaminski et al., 2008; McNeil & Fyfe, 2012); applied memory studies (Hampstead et al., 2012; Karpicke & Roediger, 2008); and even recent studies on the amelioration of heuristics and its closely related constructs (e.g. De Neys, 2021; Janssen et al., 2020; Sedlmeier & Gigerenzer, 2001; Zhu & Gigerenzer, 2006). Given the present findings and the state of the literature, I recommend that future studies explicitly consider spontaneity of transfer.

Transfer, Processes and Kinds of Memory: Relating Memory & Judgment

Reflectivity, cognitive ability, and recognition memory: who and how can one judge what one process 'truly' is and is not?[12] How many cognitive processes are there? Inspired by the numerous overlaps in activated brain areas between multiple purportedly distinct cognitive processes, Price and Friston (2005) propose the creation of a cognitive ontology: a systematic description of the cognitive processes that exist and their relationship with each other. One development of this proposal is the Cognitive Atlas headed by Russell Poldrack (Poldrack & Yarkoni, 2016). Moreover, in recorded presentations such as in Institute for Science in Society (2022), Poldrack has argued that one of the challenges in creating this map is that much of our terminology is still reminiscent of William James' introspectionist *Principles of Psychology* (James, 1890). Evidence for the importance of introspection in our theories of cognition can be found in how the predictions and interpretations in this thesis are made, the explication of the tripartite model (e.g. Stanovich et al., 2016), and even the way heuristics are conceptualized in Tversky and Kahneman's work (e.g. Kahneman, 2011). Given that introspection has historically been treated with great suspicion by psychologists (Boring, 1953; Danziger, 1980; Kukla, 2001)

[12] The material in this subsection is short adaptation from one paper submitted to Dr. Thomas Teo's PSYC6030 course in 2021 and another submitted to Dr. Ellen Bialystok's PSYC6665 course in 2022. I obtained explicit permission to re-use this submitted material and a few sentences were directly copied because I felt that there was no other way to express my argument. I am the sole author and origin of these statements.

and that introspection is fundamental to our theories, it is important to consciously consider whether we should be suspicious of our theories in the same way we are to introspection itself. For example, if we assume that the participant is not lying, there are some introspections that cannot be wrong—such as whether they feel pain or not (Waterfall, 2015). However, there are also cases where it is unclear whether the introspection given in response to a prompt correctly answers that prompt or some other question—i.e., a person's retelling of how painful an event is can differ wildly depending on whether it is an ongoing experience or a memory, yet the prompt to elicit the introspection is (essentially) identical (Redelmeier & Kahneman, 1996). In other words, the degree to which we can be confident what question an introspective report answers can vary dramatically and may depend on the state of our own psychological theories.

I propose that one potential avenue to address this state of affairs is through the study of transfer. The fundamental auxiliary hypothesis implied in most experiments that purport to test transfer is that task *A1* and *A2* share a common process *A*. Hence, typically, the argument is that if transfer exists then training on task *A1* should lead to benefits on task *A2* due to improvements on their shared process *A*. However, few papers[13] question the assumption that process *A* 'truly does exist', that it is something more than just a useful or best-available theoretical construct. This is unfortunate because unless process *A* truly does exist, then one cannot test whether transfer exists or not. Although this may appear to be an irreconciliable weakness of transfer studies, if we switch the assumption and instead take the position that transfer truly does exist, then transfer may be used to confirm or falsify the existence of cognitive processes. Hence, perhaps transfer can be used to make a map of the different cognitive processes that exist based on the way certain kinds of learning transfer to other contexts and forms. Furthermore, if we

[13] See Bialystok and Craik (2022), Gathercole et al. (2019) and von Bastian et al. (2022) for notable examples.

assume that the neurological basis for transfer of learning is due to overlapping neural instantiations between task *A1* and *A2* (Jonides, 2004), then neuroimaging can be used to constrain which tasks will exhibit transfer (Dahlin et al., 2008; Nichols et al., 2021). Importantly, because this way of confirming and falsifying the existence of processes is based on the encoding and retrieval of people's learning, then perhaps it may also address the number of memory systems that exist or whether such a question is fundamentally flawed (De Brigard et al., 2022; Renoult et al., 2019; Tulving, 2007).[14]

One immediate hurdle with this thinking is the proper interpretation of scenarios where only some of the participants demonstrate transfer of learning, as was the case here. How should we interpret this under the new proposed framework? Do different people have different cognitive processes? Memory systems? These questions are important even without discussing transfer, but the proposed view of transfer may be able to force us to consider them.

What is System 1?

Going beyond the discussed research, it is worth considering why the particular heuristics—intuitions—identified by Kahneman and Tversky appear to be the default in many people. Are these defaults acquired in the same way Stanovich et al. (2016) predict their replacement via overlearning? Are they some form of innate knowledge? Or could they be natural by-products of cognitive processing in humans, as may be implied by Gigerenzer (1991)? What is system 1 and how do our specific intuitions come to be? These questions are important because they may bear on the required intervention for overcoming our automatic intuitions. For example, if these intuitions are global by-products of cognitive processes, then it may be the case

[14] In this way, it may be the case that something like memory for past events, roughly referred to as episodic memory (Tulving, 2002), is a special case of transfer.

that training can only replace the intuitions under contexts and instances that are closely related to training. Nevertheless, the experiment discussed here concerns the dual-process theory compliant transfer success to the Linda problem and not generalized transfer success to all (problematic) instances of the representativeness heuristic and/or conjunction fallacy.

Future Directions

There are several analyses with this dataset that could shine light on the problems discussed here. Firstly, the reliability of the measures of individual-differences used here should be analyzed—such as through split-half reliability (Parsons et al., 2019; Pronk et al., 2021) and coefficient omega (Flora, 2020) —to infer how their appropriateness for individual differences analyses (Hedge et al., 2018). Secondly, participants post-experimental justifications for their responses on the Linda problem were also collected. Analyzing their justifications may provide important insights on why even explicit condition participants did not transfer their learning and also inform the debate between Panglossians and Meliorists (Gigerenzer, 1991, 1996; Stanovich, 2009; Tversky & Kahneman, 1996).

In terms of future experiments directly building on this work, I propose that efforts should be made to improve the training material. Furthermore, these findings should be conceptually replicated by introducing even larger temporal gaps between training to inhibit heuristics and the tests of successful training-related inhibition, to see how forgetting may impact inhibition of heuristics.

Conclusion

This thesis sought to fill a gap in the prior literature on learning to inhibit our heuristics and biased judgments by testing whether transfer success under dual process theory occurs or

not. I successfully replicated and extended a prior study on the problem by controlling for individual differences and masking the relationship between the learning material and its targeted problem. Using this design, I found transfer success and concluded that education truly can inhibit heuristics. Furthermore, I made several specific predictions regarding the relationships between individual differences and the different training conditions, but none of these relationships were consistently supported by my analysis of the data. However, I successfully replicated prior findings regarding the importance of reflectivity and cognitive ability to heuristics and biases, and found that accurate recognition memory is also a significant predictor even after controlling for these established individual differences. Also, I found a significant improvement in transfer with an explicit prompt to apply learning compared to its absence. Lastly, I found that the relationship between individual differences, learning materials with graphs & text versus text-only, and transfer success may vary based on this prompting. This variation in the three-way relationship means that if the natural context differs in the spontaneity 'required' for transfer compared to the experimental context, then the studied relationship between individual differences and the specific learning material may be incorrect. Therefore, future studies should address the problem of spontaneity not just to improve the accuracy of their estimates of transfer success and validity, but also to achieve a more complete picture of the roles of individual differences and training types in their studies.

Postface

Is everything a memory? The design and interpretation of this experiment was motivated by an attempt to theoretically unify the fields of memory and judgment. I did this in following recent trends in memory research wherein an increasing number of previously independent cognitive functions are now being subsumed under its purview (Biderman et al., 2020; Cabeza et al., 2020; Danziger, 2008; Klein, 2015; Madore et al., 2015; Moscovitch et al., 2016; Schacter et al., 2012; Thakral et al., 2020). Of the challenges to this claim that everything is a memory—sometimes referred to as associationism—the one that I focused on here is the one of 'productivity': that an associationistic theory of mind has difficulty explaining how people's thoughts can be combined into an infinite number of different meaningful variations (Fodor & Pylyshyn, 1988). I argue that this challenge can be restated in terms of the problem of 'transfer', of whether learning—prior memories—can be transferred from one form and context to another and be intelligently applied (Barnett & Ceci, 2002). Both productivity and transfer point to the same underlying issue: if memory is so central to cognition, then how can we reconcile its—traditionally perceived—limited and passé nature with acts of creativity and ingenuity? Although, restating the challenge of productivity in terms of the transfer of learned behaviors makes the problem more tangible, it is also problematic in that the evidence for transfer is so controversial (Barnett & Ceci, 2002; Kahneman & Klein, 2009; Nichols et al., 2021; Owen et al., 2010; Simons et al., 2016; von Bastian et al., 2022). Nevertheless and at the very least, this controversy needs to be addressed because of what it may imply regarding education.

References

Agnoli, F. (1991). Development of judgmental heuristics and logical reasoning: Training counteracts the representativeness heuristic. *Cognitive Development*, *6*(2), 195–217. https://doi.org/10.1016/0885-2014(91)90036-D

Agnoli, F., & Krantz, D. H. (1989). Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy. *Cognitive Psychology*, *21*(4), 515–550. https://doi.org/10.1016/0010-0285(89)90017-0

Agresti, A. (2019). *An introduction to categorical data analysis* (Third edition). John Wiley & Sons.

Arel-Bundock, V. (2022). *marginaleffects: Marginal Effects, Marginal Means, Predictions, and Contrasts* (R package version 0.6.0.9000). https://vincentarelbundock.github.io/marginaleffects/

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Barnett, S. M., & Ceci, S. J. (2005). Reframing the evaluation of education: Assessing whether learning transfers beyond the classroom. In *Transfer of Learning from a Multidisciplinary Perspective. Greenwich, Connecticut: Information Age Publishing*.

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, *50*(5), 1953–1959. https://doi.org/10.3758/s13428-017-0963-x

Bialystok, E. (2017). The bilingual adaptation: How minds accommodate experience. *Psychological Bulletin*, *143*(3), 233–262. https://doi.org/10.1037/bul0000099

Bialystok, E. (2021). Bilingualism: Pathway to Cognitive Reserve. *Trends in Cognitive Sciences*, *25*(5), 355–364. https://doi.org/10.1016/j.tics.2021.02.003

Bialystok, E., & Craik, F. I. M. (2022). How does bilingualism modify cognitive function? Attention to the mechanism. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-022-02057-5

Biderman, N., Bakkour, A., & Shohamy, D. (2020). What Are Memories For? The Hippocampus Bridges Past Experience with Future Decisions. *Trends in Cognitive Sciences*, *24*(7), 542–556. https://doi.org/10.1016/j.tics.2020.04.004

Bock, D. D., Deprez, J., Dooren, W. V., Roelens, M., & Verschaffel, L. (2011). Abstract or Concrete Examples in Learning Mathematics? A Replication and Elaboration of Kaminski, Sloutsky, and Heckler's Study. *Journal for Research in Mathematics Education*, *42*(2), 109–126. https://doi.org/10.5951/jresematheduc.42.2.0109

Boring, E. G. (1953). A history of introspection. *Psychological Bulletin*, *50*(3), 169–189. https://doi.org/10.1037/h0090793

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Cabeza, R., Becker, M., & Davis, S. W. (2020). Are the hippocampus and its network necessary for creativity? *Proceedings of the National Academy of Sciences*, 202008601. https://doi.org/10.1073/pnas.2008601117

Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Clark, I. A., Kim, M., & Maguire, E. A. (2018). Verbal Paired Associates and the Hippocampus: The Role of Scenes. *Journal of Cognitive Neuroscience*, *30*(12), 1821–1845. https://doi.org/10.1162/jocn_a_01315

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of Learning after Updating Training Mediated by the Striatum. *Science*, *320*(5882), 1510–1512.

Danziger, K. (1980). The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences*, *16*(3), 241–262. https://doi.org/10.1002/1520-6696(198007)16:3<241::AID-JHBS2300160306>3.0.CO;2-O

Danziger, K. (2008). *Marking the mind: A history of memory*. Cambridge University Press.

De Brigard, F., Umanath, S., & Irish, M. (2022). Rethinking the distinction between episodic and semantic memory: Insights from the past, present, and future. *Memory & Cognition*, *50*(3), 459–463. https://doi.org/10.3758/s13421-022-01299-x

De Neys, W. (2021). On Dual- and Single-Process Models of Thinking. *Perspectives on Psychological Science*, 1745691620964172. https://doi.org/10.1177/1745691620964172

Evans, J. St. B. T., Handley, S. J., Neilens, H., & Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *Quarterly Journal of Experimental Psychology*, *63*(5), 892–909. https://doi.org/10.1080/17470210903111821

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, *50*(2), 123–129. https://doi.org/10.1007/BF00309212

Fisk, J. E., & Pidgeon, N. (1997). The conjunction fallacy: The case for the existence of competing heuristic strategies. *British Journal of Psychology*, *88*(1), 1–27. https://doi.org/10.1111/j.2044-8295.1997.tb02617.x

Flora, D. B. (2020). Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Fox, J. (2016). *Applied regression analysis and generalized linear models* (Third Edition). SAGE.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., … R-Core. (2022). *car: Companion to Applied Regression* (3.1-0). https://CRAN.R-project.org/package=car

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Friendly, M., & Meyer, D. (2015). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b19022

Friendly, M., Turner, H., Zeileis, A., Murdoch, D., Firth, D., Kumar, M., & Sun, S. (2022). *vcdExtra: "vcd" Extensions and Additions* (0.8-0). https://CRAN.R-project.org/package=vcdExtra

Fyfe, E. R., McNeil, N. M., & Borjas, S. (2015). Benefits of "concreteness fading" for children's mathematics understanding. *Learning and Instruction*, *35*, 104–120. https://doi.org/10.1016/j.learninstruc.2014.10.004

Gathercole, S. E., Dunning, D. L., Holmes, J., & Norris, D. (2019). Working memory training involves learning new skills. *Journal of Memory and Language*, *105*, 19–42. https://doi.org/10.1016/j.jml.2018.10.003

Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases." *European Review of Social Psychology*, *2*(1), 83–115. https://doi.org/10.1080/14792779143000033

Gigerenzer, G. (1996). *On narrow norms and vague heuristics: A reply to Kahneman and Tversky.* - *PsycNET*. https://doi.apa.org/doiLanding?doi=10.1037%2F0033-295X.103.3.592

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gohel, D., Jager, C., Fazilleau, Q., output), M. N. (rmarkdown for docx, Robert, T., footnotes), M. B. (inline, reference), A. Y. (support for bookdown cross, objects), P. J. (support for gam, & system), S. B. (work on footnote positioning. (2022). *flextable: Functions for Tabular Reporting* (0.7.2). https://CRAN.R-project.org/package=flextable

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90. https://doi.org/10.1037/0033-295X.109.1.75

Goldstone, R. L., & Son, J. Y. (2005). The Transfer of Scientific Principles Using Concrete and Idealized Simulations. *Journal of the Learning Sciences*, *14*(1), 69–110. https://doi.org/10.1207/s15327809jls1401_4

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as Effect Sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443–455. https://doi.org/10.1177/2515245920922982

Hagenaars, S. P., Harris, S. E., Davies, G., Hill, W. D., Liewald, D. C. M., Ritchie, S. J., Marioni, R. E., Fawns-Ritchie, C., Cullen, B., Malik, R., Worrall, B. B., Sudlow, C. L. M., Wardlaw, J. M., Gallacher, J., Pell, J., McIntosh, A. M., Smith, D. J., Gale, C. R., & Deary, I. J. (2016). Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Molecular Psychiatry*, *21*(11), 1624–1632. https://doi.org/10.1038/mp.2015.225

Hampstead, B. M., Stringer, A. Y., Stilla, R. F., Giddens, M., & Sathian, K. (2012). Mnemonic strategy training partially restores hippocampal activity in patients with mild cognitive impairment. *Hippocampus*, *22*(8), 1652–1658. https://doi.org/10.1002/hipo.22006

Hanmer, M. J., & Ozan Kalkan, K. (2013). Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models. *American Journal of Political Science*, *57*(1), 263–277. https://doi.org/10.1111/j.1540-5907.2012.00602.x

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Heiss, A. (2022, May 20). Marginalia: A guide to figuring out what the heck marginal effects, marginal slopes, average marginal effects, marginal effects at the mean, and all these other marginal things are. *Andrew Heiss*. https://www.andrewheiss.com/blog/2022/05/20/marginalia/

Hertwig, R., Benz, B., & Krauss, S. (2008). The conjunction fallacy and the many meanings of and. *Cognition*, *108*(3), 740–753. https://doi.org/10.1016/j.cognition.2008.06.008

Hertwig, R., & Chase, V. M. (1998). Many Reasons or Just One: How Response Mode Affects Reasoning in the Conjunction Problem. *Thinking & Reasoning*, *4*(4), 319–352. https://doi.org/10.1080/135467898394102

Hertwig, R., & Gigerenzer, G. (1999). The `conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*(4), 31.

Iannone, R. (2020). *DiagrammeR: Graph/Network Visualization* (1.0.6.1). https://CRAN.R-project.org/package=DiagrammeR

Institute for Science in Society. (2022, January 24). *Russell A. Poldrack "(How) can neuroimaging inform the architecture of the mind?" (#DDLS).* https://www.youtube.com/watch?v=8oaJ21kSL-4

James, W. (1890). The Principles of Psychology, New York: Henry Holt. *Reprinted Cambridge, MA: Harvard*.

Janssen, E. M., Raoelison, M., & de Neys, W. (2020). "You're wrong!": The impact of accuracy feedback on the bat-and-ball problem. *Acta Psychologica*, *206*, 103042. https://doi.org/10.1016/j.actpsy.2020.103042

Jones, M. G. (2009a). Research Commentary: Examining Surface Features in Context. *Journal for Research in Mathematics Education*, *40*(2), 94–96. https://doi.org/10.2307/40539327

Jones, M. G. (2009b). Research Commentary: Transfer, Abstraction, and Context. *Journal for Research in Mathematics Education*, *40*(2), 80–89. https://doi.org/10.5951/jresematheduc.40.2.0080

Jonides, J. (2004). How does practice makes perfect? *Nature Neuroscience*, *7*(1), 10–11. https://doi.org/10.1038/nn0104-10

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed). Farrar, Straus and Giroux.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The Advantage of Abstract Examples in Learning Math. *Science*, *320*(5875), 454–455. https://doi.org/10.1126/science.1154659

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2013). The cost of concreteness: The effect of nonessential information on analogical transfer. *Journal of Experimental Psychology: Applied*, *19*(1), 14–29. https://doi.org/10.1037/a0031931

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629. https://doi.org/10.1017/psrm.2020.6

Kent, P. L. (2017). Evolution of Wechsler's Memory Scales: Content and structural analysis. *Applied Neuropsychology: Adult*, *24*(3), 232–251. https://doi.org/10.1080/23279095.2015.1135798

Klein, S. B. (2015). What memory is. *WIREs Cognitive Science*, *6*(1), 1–38. https://doi.org/10.1002/wcs.1333

Kubic, M. (2016). *CommonLit | The Roaring Twenties*. CommonLit. https://www.commonlit.org/en/texts/the-roaring-twenties

Kukla, A. (2001). *Methods of theoretical psychology*. MIT Press. http://cognet.mit.edu/library/books/view?isbn=0262112612

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Larsson, J., & Gustafsson, P. (2018). A Case Study in Fitting Area-Proportional Euler Diagrams with Ellipses Using eulerr. *Proceedings of International Workshop on Set Visualization and Reasoning*, *2116*, 84–91. https://cran.r-project.org/package=eulerr

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (1.7.5). https://CRAN.R-project.org/package=emmeans

Lewis, M. (2017). *The undoing project: A friendship that changed our minds* (First edition). W.W. Norton & Company.

Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772

Lyall, D. M., Cullen, B., Allerhand, M., Smith, D. J., Mackay, D., Evans, J., Anderson, J., Fawns-Ritchie, C., McIntosh, A. M., Deary, I. J., & Pell, J. P. (2016). Cognitive Test Scores in UK Biobank: Data Reduction in 480,416 Participants and Longitudinal Stability in 20,346 Participants. *PLOS ONE*, *11*(4), e0154222. https://doi.org/10.1371/journal.pone.0154222

Madore, K. P., Addis, D. R., & Schacter, D. L. (2015). Creativity and Memory: Effects of an Episodic-Specificity Induction on Divergent Thinking. *Psychological Science*, *26*(9), 1461–1468. https://doi.org/10.1177/0956797615591863

Makowski, D. (2018). The psycho Package: An Efficient and Publishing-Oriented Workflow for Psychological Science. *Journal of Open Source Software*, *3*(22), 470. https://doi.org/10.21105/joss.00470

McNeil, N. M., & Fyfe, E. R. (2012). "Concreteness fading" promotes transfer of mathematical knowledge. *Learning and Instruction*, *22*(6), 440–448. https://doi.org/10.1016/j.learninstruc.2012.05.001

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration. *Psychological Science*, *12*(4), 269–275. https://doi.org/10.1111/1467-9280.00350

Meltzer, J. A., Kates Rose, M., Le, A. Y., Spencer, K. A., Goldstein, L., Gubanova, A., Lai, A. C., Yossofzai, M., Armstrong, S. E. M., & Bialystok, E. (2021). Improvement in executive function for older adults through smartphone apps: A randomized clinical trial comparing language learning and brain training. *Aging, Neuropsychology, and Cognition*, 1–22. https://doi.org/10.1080/13825585.2021.1991262

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making*, *13*(3), 246–259.

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic Memory and Beyond: The Hippocampus and Neocortex in Transformation. *Annual Review of Psychology*, *67*(1), 105–134. https://doi.org/10.1146/annurev-psych-113011-143733

Nichols, E. S., Erez, J., Stojanoski, B., Lyons, K. M., Witt, S. T., Mace, C. A., Khalid, S., & Owen, A. M. (2021). Longitudinal white matter changes associated with cognitive training. *Human Brain Mapping*, *42*(14), 4722–4739. https://doi.org/10.1002/hbm.25580

Nichols, E. S., Wild, C. J., Stojanoski, B., Battista, M. E., & Owen, A. M. (2020). Bilingualism Affords No General Cognitive Advantages: A Population Study of Executive Function in 11,000 People. *Psychological Science*, *31*(5), 548–567. https://doi.org/10.1177/0956797620903113

Olszewska, A. M., Gaca, M., Herman, A. M., Jednoróg, K., & Marchewka, A. (2021). How Musical Training Shapes the Adult Brain: Predispositions and Neuroplasticity. *Frontiers in Neuroscience*, *15*, 630829. https://doi.org/10.3389/fnins.2021.630829

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R. J., & Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*(7299), 775–778. https://doi.org/10.1038/nature09042

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, *4*(1), 32–38. https://doi.org/10.1016/S0022-5371(65)80064-0

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning Styles: Concepts and Evidence. *Psychological Science in the Public Interest*, *9*(3), 105–119. https://doi.org/10.1111/j.1539-6053.2009.01038.x

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Poldrack, R. A., & Yarkoni, T. (2016). From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. *Annual Review of Psychology*, *67*(1), 587–612. https://doi.org/10.1146/annurev-psych-122414-033729

Price, C. J., & Friston, K. J. (2005). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology*, *22*(3–4), 262–275. https://doi.org/10.1080/02643290442000095

Prolific (March to May 2022). (2022). Prolific. https://www.prolific.co/

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2021). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-01948-3

Qualtrics (March to May 2022). (2022). Qualtrics. https://www.qualtrics.com/

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*(1), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6

Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From Knowing to Remembering: The Semantic-Episodic Distinction. *Trends in Cognitive Sciences*, *23*(12), 1041–1057. https://doi.org/10.1016/j.tics.2019.09.008

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The Future of Memory: Remembering, Imagining, and the Brain. *Neuron*, *76*(4), 677–694. https://doi.org/10.1016/j.neuron.2012.11.001

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Toomet, O., Crowley, J., Hofmann, H., & Wickham, H. (2021). *GGally: Extension to "ggplot2"* (2.1.2). https://CRAN.R-project.org/package=GGally

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380–400. https://doi.org/10.1037/0096-3445.130.3.380

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do "Brain-Training" Programs Work? *Psychological Science in the Public Interest*, *17*(3), 103–186. https://doi.org/10.1177/1529100616661983

Smart, J., Sutherland, W. J., Watkinson, A. R., & Gill, J. A. (2004). A New Means of Presenting the Results of Logistic Regression. *The Bulletin of the Ecological Society of America*, *85*(3), 100–102. https://doi.org/10.1890/0012-9623(2004)85[100:ANMOPT]2.0.CO;2

Snodgrass, J. G., & Corwin, J. (1988). *Pragmatics of Measuring Recognition Memory: Applications to Dementia and Amnesia*. 17.

Stanovich, K. E. (2000). Individual differences in reasoning: Implications for the rationality debate? *BEHAVIORAL AND BRAIN SCIENCES*, 82.

Stanovich, K. E. (2009). How Bad Is Our Decision Making: The Great Rationality Debate. In *Decision Making and Rationality in the Modern World* (p. 33).

Stanovich, K. E. (2016). The Comprehensive Assessment of Rational Thinking. *Educational Psychologist*, *51*(1), 23–34. https://doi.org/10.1080/00461520.2015.1125787

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & Toplak, M. E. (2016a). Overcoming Miserly Processing: Detection, Override, and Mindware. In *The rationality quotient: Toward a test of rational thinking* (pp. 39–62). MIT Press.

Stanovich, K. E., & Toplak, M. E. (2016b). Probabilistic and Statistical Reasoning. In *The rationality quotient: Toward a test of rational thinking*. MIT Press.

Stanovich, K. E., & Toplak, M. E. (2016c). Rationality, Intelligence, and the Functional Architecture of the Mind. In *The rationality quotient: Toward a test of rational thinking* (pp. 15–38). MIT Press.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. MIT Press.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning*, *23*(3), 207–234. https://doi.org/10.1080/13546783.2017.1292954

Thakral, P. P., Madore, K. P., Kalinowski, S. E., & Schacter, D. L. (2020). Modulation of hippocampal brain networks produces changes in episodic simulation and divergent thinking. *Proceedings of the National Academy of Sciences*, 202003535. https://doi.org/10.1073/pnas.2003535117

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, *53*(1), 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Tulving, E. (2007). Are There 256 Different Kinds of Memory. *The Foundations of Remembering: Essays in Honor of Henry L. Roediger, III*, 39–52.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Tversky, A., & Kahneman, D. (1992). *Advances in prospect theory: Cumulative representation of uncertainty*. 27.

Tversky, A., & Kahneman, D. (1996). *On the reality of cognitive illusions. - PsycNET*. https://doi.apa.org/doiLanding?doi=10.1037%2F0033-295X.103.3.582

von Bastian, C. C., Belleville, S., Udale, R. C., Reinhartz, A., Essounni, M., & Strobach, T. (2022). Mechanisms underlying training-induced cognitive change. *Nature Reviews Psychology*, *1*(1), 30–41. https://doi.org/10.1038/s44159-021-00001-3

Ward, E. V., Berry, C. J., Shanks, D. R., Moller, P. L., & Czsiser, E. (2020). Aging Predicts Decline in Explicit and Implicit Memory: A Life-Span Study. *Psychological Science*, *31*(9), 1071–1083. https://doi.org/10.1177/0956797620927648

Waterfall, D. E. (2015). Wittgenstein on Introspection and Introspectionism. *Sophia*, *54*(3), 243–264. https://doi.org/10.1007/s11841-015-0468-y

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed. 2016). Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-24277-4

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2"* (1.1.1). https://CRAN.R-project.org/package=cowplot

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., & Wilke, C. O. (2020). colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. *Journal of Statistical Software*, *96*(1), 1–49. https://doi.org/10.18637/jss.v096.i01

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*(3), 287–308. https://doi.org/10.1016/j.cognition.2004.12.003

Appendices

Appendix A: Regarding The Use of the Linda Problem and a Single vs. Multiple Problems

One potential drawback of this design (Figure 1) is that it becomes problematic to introduce multiple CE problems, which is unfortunate because having multiple CE problems might allow the use of parametric tests that rely on a normally distributed continuous response variable. However, I argue that this is not a significant problem and that having multiple CE problems would introduce various risks, including potentially endangering how well dual-process compliant transfer success is tested in the implicit condition.

Firstly, each additional CE problem other than the Linda problem increases the risk of confounding our results and would make it difficult to claim that we have made significant progress from Agnoli and Krantz (1989) because they may clue-in participants to the purpose of the experiment. Even if we were to accept this risk, the problem becomes how many questions are necessary to approximate a normally distributed continuous response variable. I have been unsuccessful in finding a definitive answer to this problem, but I assume it has to be at least three. Having three conjunction problems in the heuristics and biases battery would mean that they make up about 3/17 = 17.6% of the battery, which intuitively sounds like substantial opportunity for a participant to realize the scope of the experiment.

Secondly, to my knowledge, the only other CE problem that has been shown to have comparable resistance to transfer—where statistically-sophisticated and naïve participants perform similarly—is the much less discussed Bill problem (Tversky & Kahneman, 1983). Previously shown transfer resistance is important, otherwise one could criticize the experiment for using problems for which transfer is easier, in order to inflate the estimate of the effect that education has on inhibiting heuristics. Furthermore, although the Linda problem is highly

controversial, this does not mean that using other conjunction-type problems grants immunity to the same criticisms directed towards the Linda problem. To my knowledge, other conjunction-type problems have not received the same attention, presumably because they are simply not as notorious.

Thirdly, although it has been kindly suggested by a statistical consultant that one may use a multivariate analysis to quantify the effect that the other CE problems have on cuing transfer to the Linda problem—presumably by analyzing whether correct Linda problem responses co-occur with other correct CE problem-responses exclusively or not—I argue that even multivariate analysis may fail here. The problem is that it is conceivable that regardless of whether a participant correctly answers a previously shown CE problem or not, if they realize that the Linda problem bears resemblance to a previous CE problem and hence the true aim of the learning materials, then using their response to the Linda problem as a measure of transfer success—as it is described here—is invalidated. Thus, because responses on other CE problems are uninformative of whether they cue correct responses on the Linda problem or not—by hinting at the intent of the experiment—I suspect that even multivariate analysis cannot properly quantify the effect that other CE problems have on the Linda problem.

Overall, taking these risks into account, and given that using univariate binomial regression—specifically logistic regression—has already been shown to be appropriate for our situation (Fox, 2016), I argue it was best to use only the Linda problem.

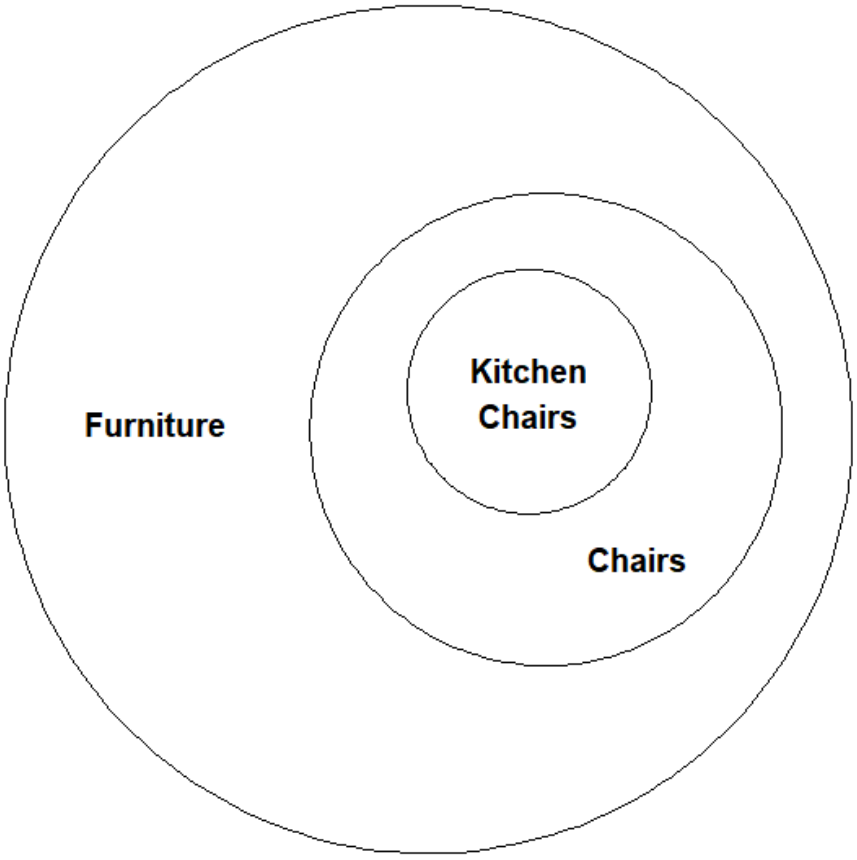
Appendix B: Learning Material Examples

Below is an adaptation of the training module provided in Appendix A of Agnoli and Krantz (1989). Figures and Venn diagrams were modernized to increase conceptual consistency and prevent misunderstandings in the subsection *With Venn Diagrams*. Diagrams were programmatically generated using the R package eulerr (Larsson & Gustafsson, 2018). Some text was removed or modified to: (1) ensure that the purpose of the training was masked in the implicit condition; and (2) remove any mention or reference to diagrams in the text-only learning material condition.

With Venn Diagrams

In this section, we are interested in studying how people reason and the kinds of difficulties they often encounter in solving problems. By studying the difficulties that people encounter in reasoning, we may be able to find general rules by which the human mind works.

Specifically, we are interested in studying how well people are able to learn how to reason about the frequencies of different categories or about the relationships between categories. One tool that helps people think about categories is diagrams. For example, here is a diagram that shows the relationship among three categories: (1) furniture, (2) chairs, and (3) kitchen chairs. You already know that "all kitchen chairs are chairs" and that "all chairs are a type of furniture." We may represent this relationship of inclusion among the three categories with three nested circles:



As you can see from the figure, the circle representing the category of kitchen chairs must be smaller than that representing the category of chairs, because we know that there are other types of chairs (for example, desk chairs and highchairs). Similarly, the circle representing chairs must be of a smaller size than that representing furniture. There are, in fact, other types of furniture besides chairs (for example, tables, sofa, and desks).

[NEW PAGE]

To see if you are able to use these nested diagrams, consider how you would use them to show the relationship among the following three categories: (1) vehicles, (2) cars, and (3) Toyotas. Then, answer the following question.

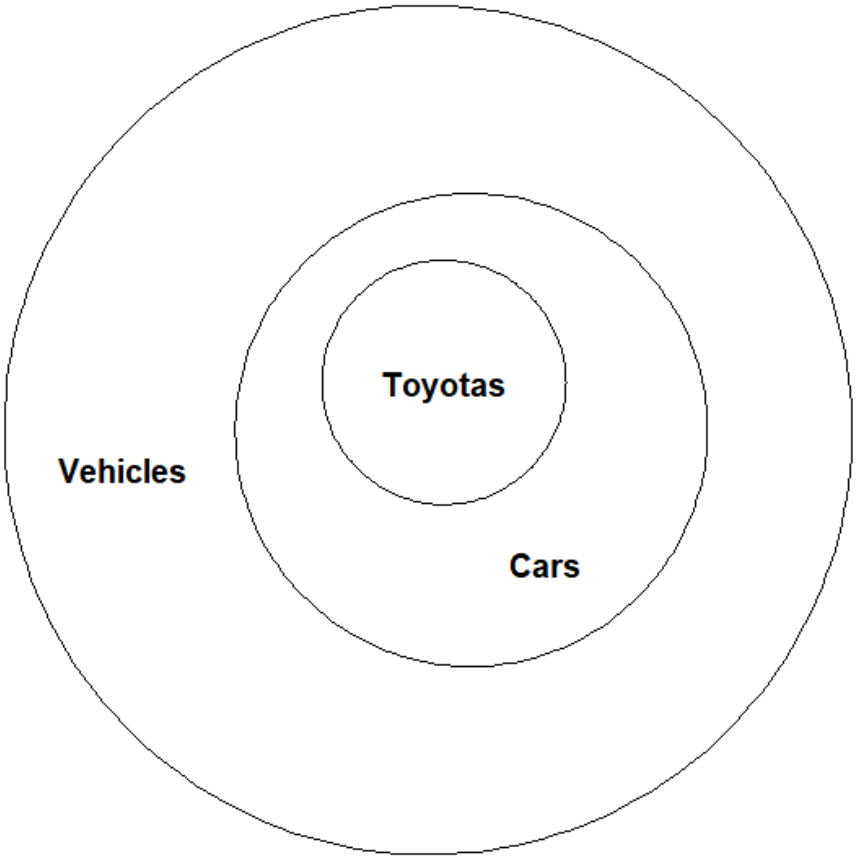
Use your mouse to rank the following categories from smallest to largest:

___ Vehicles

___ Cars

___ Toyotas

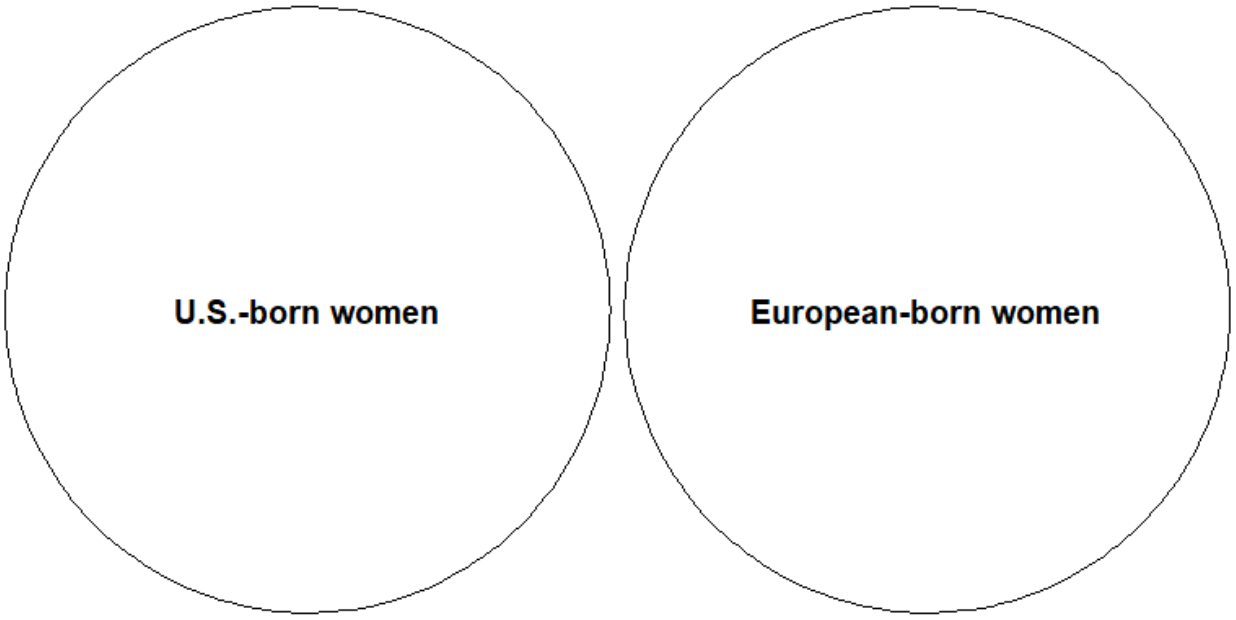
Participant receives feedback via diagram and answer response on a new page.



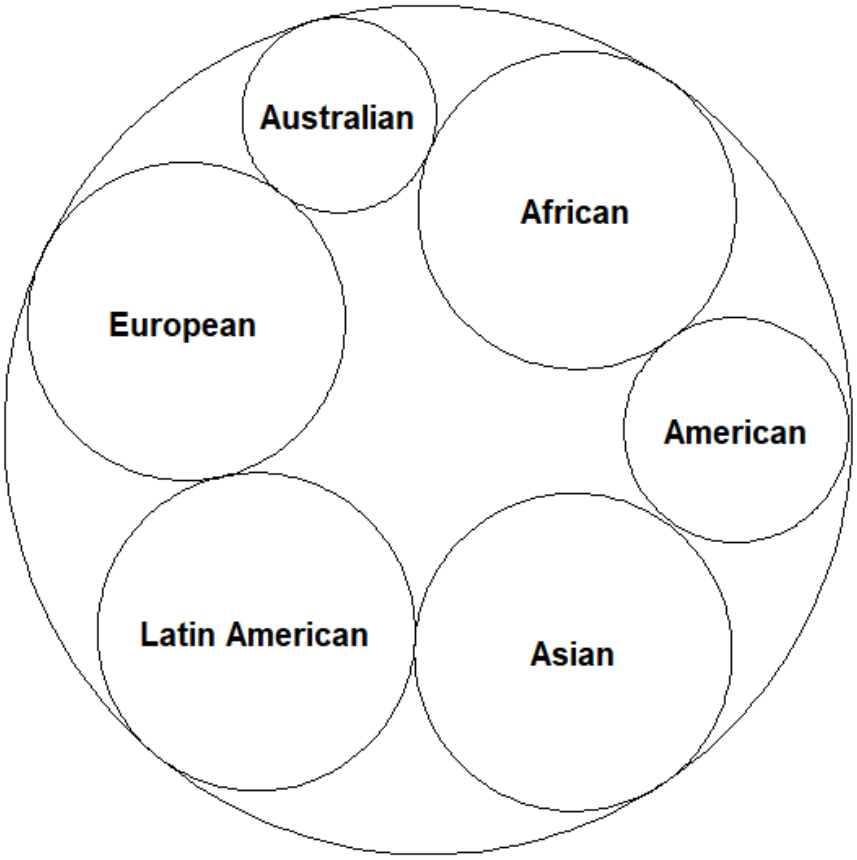
The vehicles category is larger than the cars category, which is larger than the Toyotas category.

[NEW PAGE]

The diagrams that you have seen so far represented categories that are nested within each other. Sometimes, however, categories are completely non-overlapping. Think about these two categories: (1) U.S.-born women and (2) European-born women. Membership in one of these two categories excludes the possibility of belonging to the other category. Let us represent these two categories with two circles that do not overlap at all:



Another way to represent these two categories is by dividing one figure representing all women into separate sections for women born in different parts of the world. These sections do not overlap.

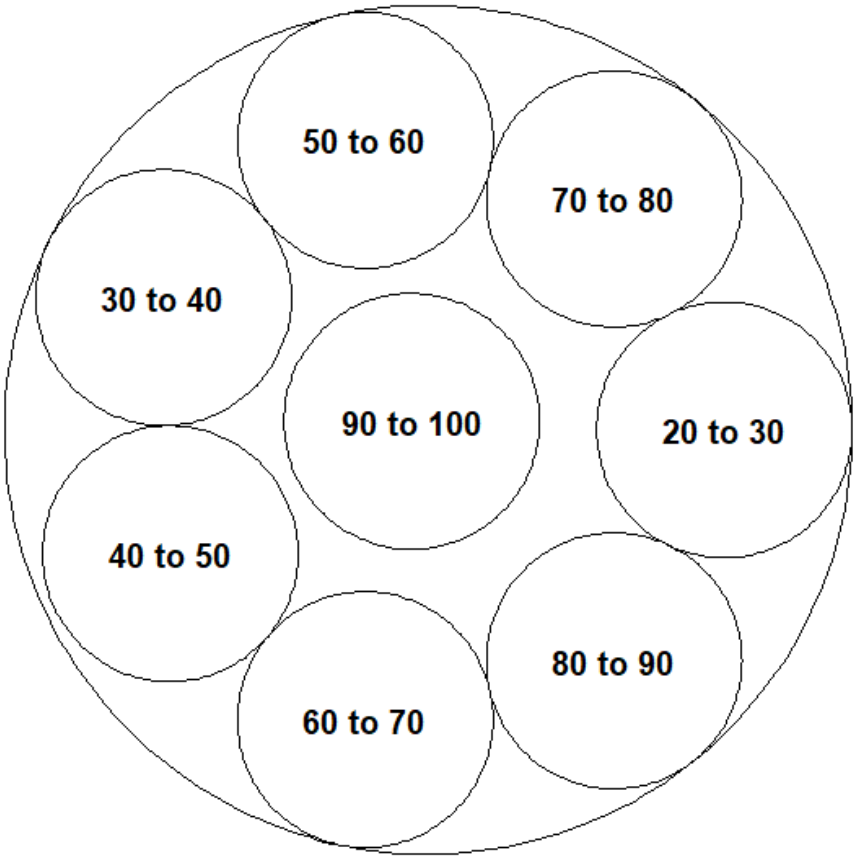


As you can see, most of the subparts of the entire figure are approximately the same size. You may know that there are many more Asian-born than Australian-born women. At times, you may want to represent this fact by assigning a larger size subpart to the group with higher frequencies and a smaller size subpart to the group with lower frequencies. Other times, you don't have to think that hard about the sizes and you can just represent the inclusion or exclusion relations by the nesting or non-overlapping of the parts.

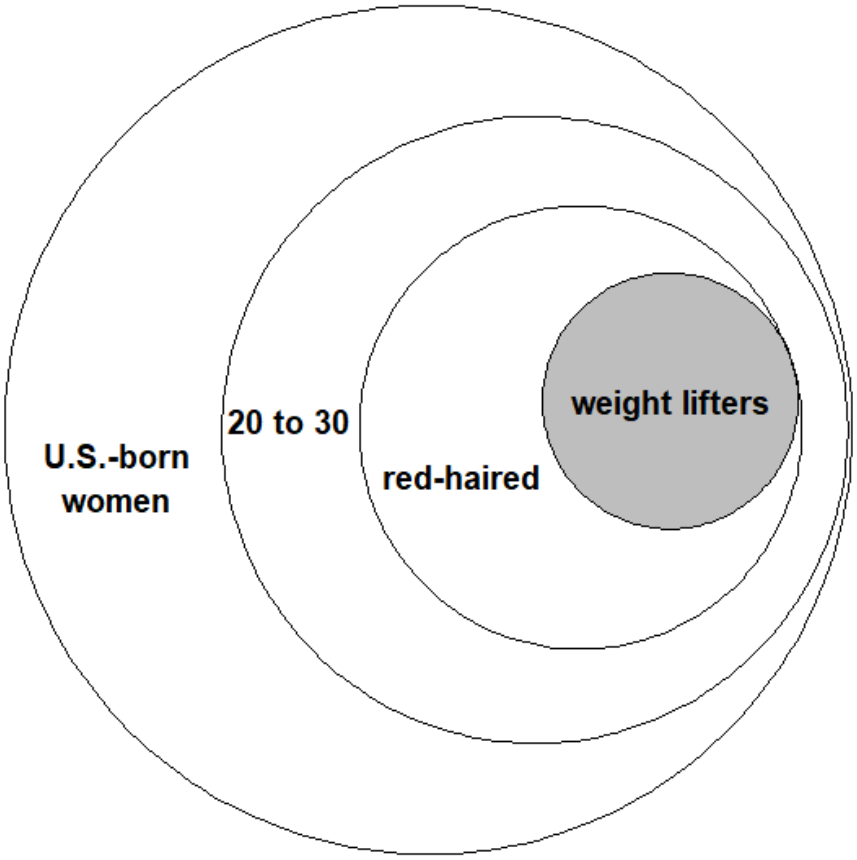
[NEW PAGE]

Now let's take one of these categories and see how we can segment it. If we take the category of U.S.-born women, we can segment the entire category by age, creating, for example,

eight subcategories of women who are 20 to 30, 30 to 40, 40 to 50, 50 to 60, 60 to 70, 70 to 80, 80 to 90, or 90 to 100 years of age. We can represent this subdivision as follows:



Now we can take the subgroup of U.S.-born women, aged 20 to 30 and think of another subgroup nested within it. For example, we can divide this group by hair color. An appropriate subgroup would be all those U.S.-born women, aged 20 to 30 who have red hair. We can continue this process by adding another characteristic: U.S.-born women, aged 20 to 30, red-haired, and weightlifters. These subsequent subdivisions can be represented as follows:



The shaded area represents the small subset of U.S.-born women who are 20 to 30 years of age and red-haired and weightlifters.

Sometimes, in order to make correct inferences it is important to accurately compare the size of two subgroups. If we ask you, for example, which of these two subgroups is larger,

Subgroup 1, U.S.-born women, 20 to 30 years of age
Subgroup 2, U.S.-born women, 20 to 30 years of age and red-haired,

you can use the graphical representation that has been illustrated to recognize that subgroup 1 must be larger than subgroup 2. Remember that it does not matter whether you use circles or any other shapes or segments of shapes to represent relationships between categories. Sometimes one kind of shape seems more convenient or just more pleasing to contemplate.

[NEW PAGE]

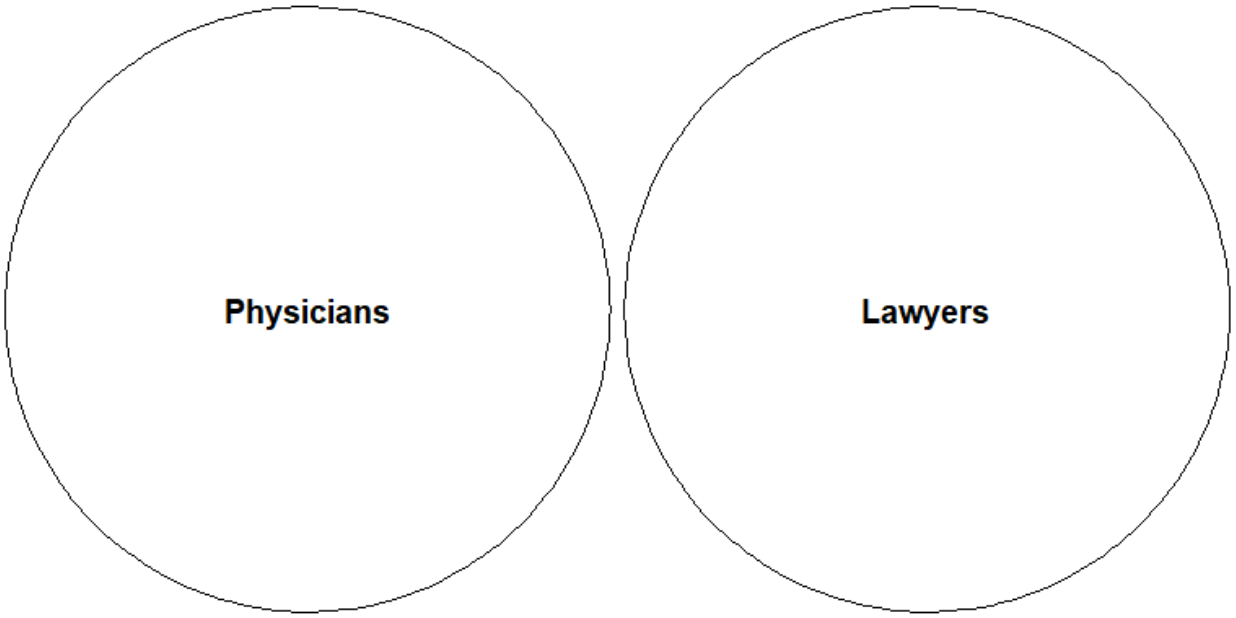
Let’s try to practice what you have been learning.

Consider how you would represent the following categories: (1) physicians and (2) lawyers.

For simplicity, let’s say that these two categories are completely non-overlapping.

Click >> for the solution

[NEW PAGE]



[NEW PAGE]

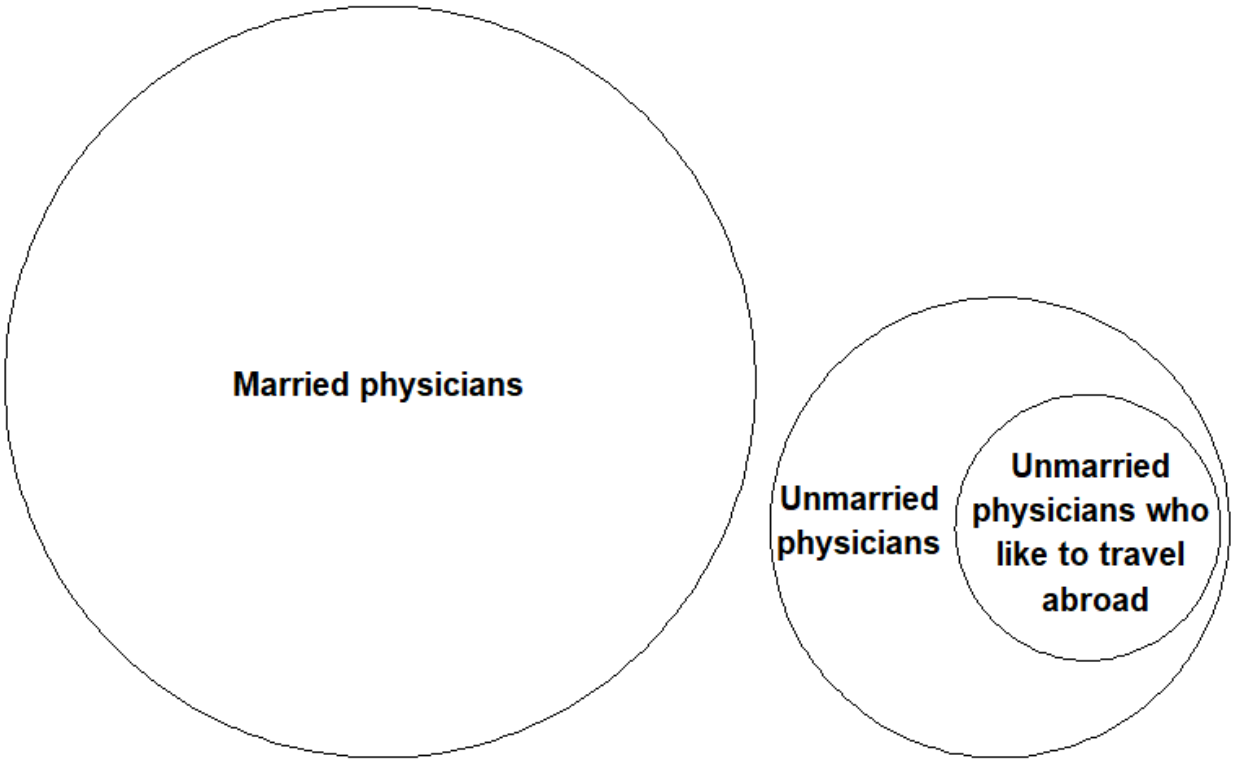
Now pick the category of physicians and consider how you would subdivide it into the two subcategories of married and unmarried physicians. Make a subsequent distinction within the subcategory of unmarried physicians by imagining that some like to travel abroad, and some don't.

Now, of the following two subgroups, select what is larger:

___ Unmarried physicians
___ Unmarried physicians who like to travel abroad

[NEW PAGE]

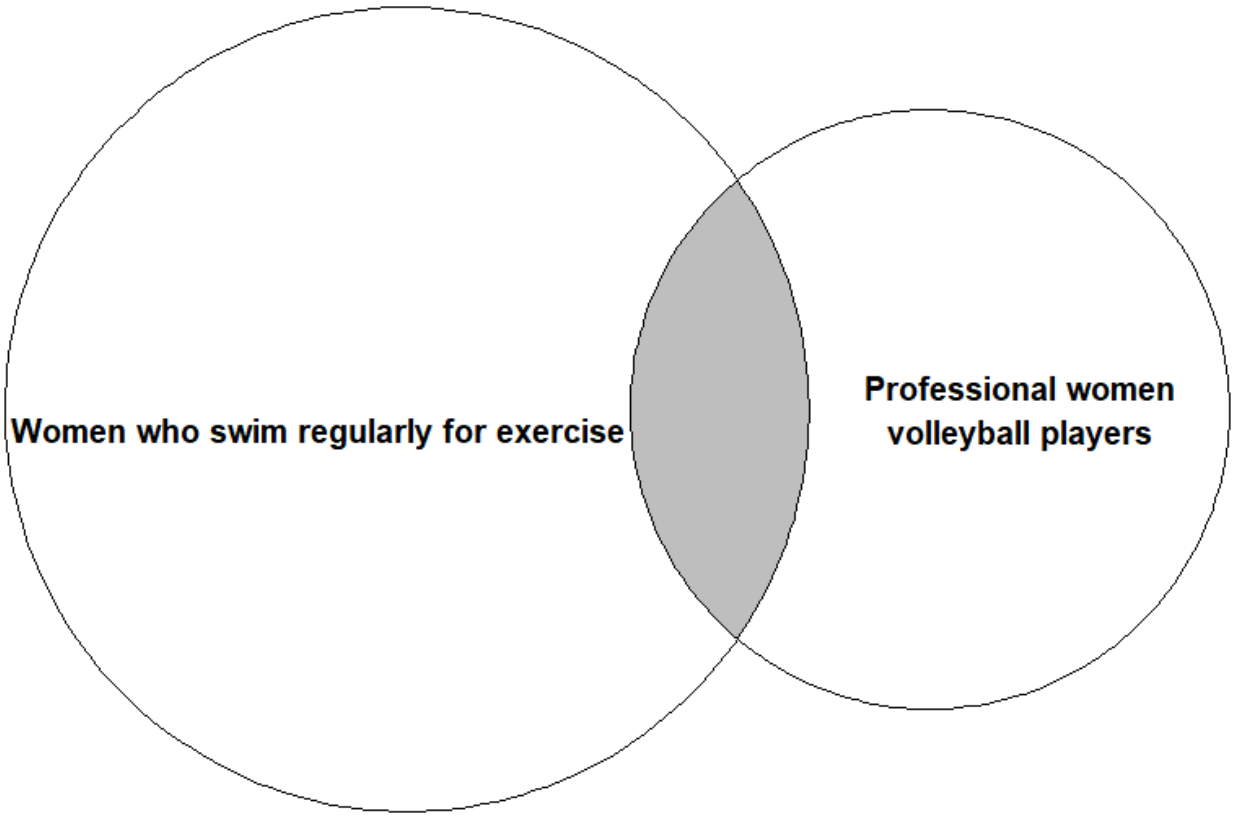
Participant receives feedback



The subgroup of unmarried physicians is larger than the subgroup of unmarried physicians who like to travel abroad.

[NEW PAGE]

Finally, we want to point out to you that some categories can partially overlap without being identical. Think about the following two categories: (1) Professional women volleyball players and (2) Women who swim regularly for exercise. The correct pictorial representation would be the following:

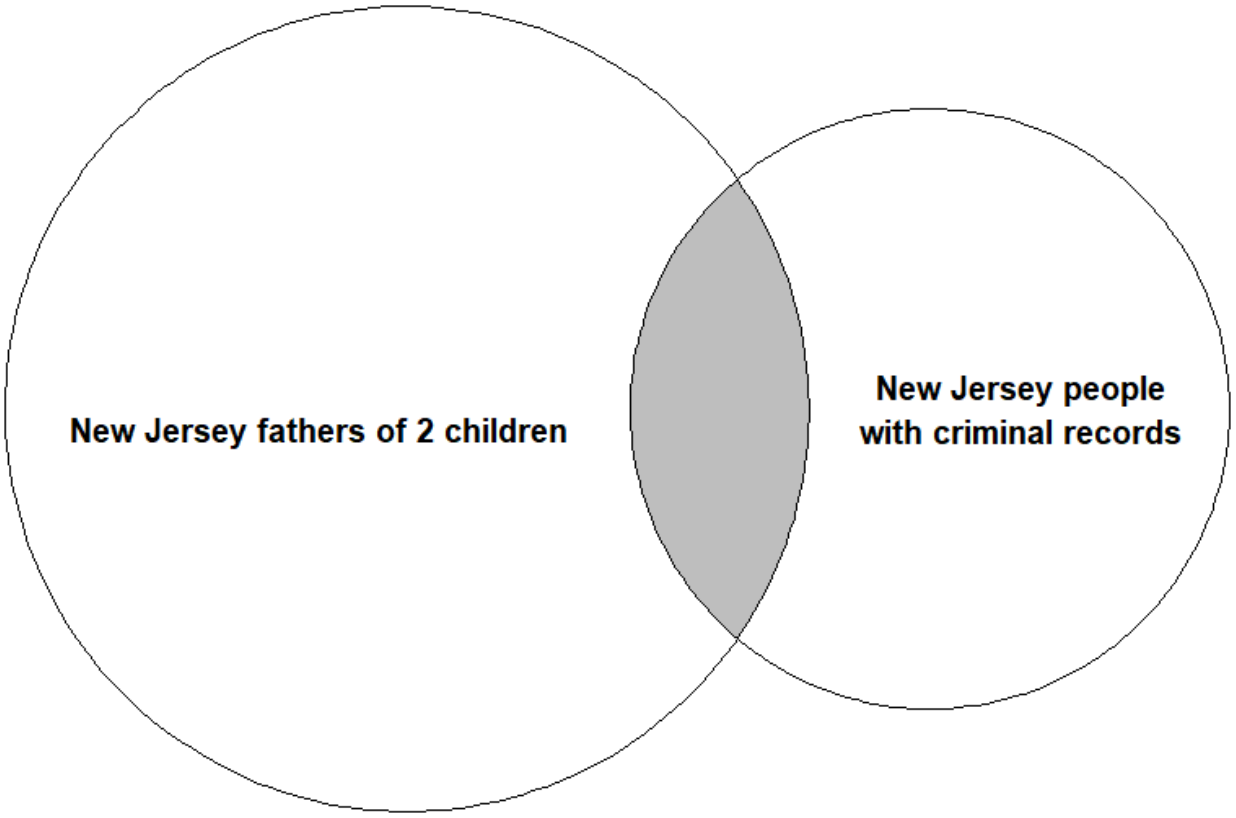


The shaded area represents those professional volleyball players who also swim regularly for exercise. As you can infer from the figure, there is a greater chance of being a woman who just swims for exercise than of being a woman who swims and is a professional volleyball

player, too. Also, there is a greater chance of being just a professional women volleyball player than of being a volleyball player who also swims regularly for exercise. Do you see that the shaded area must be smaller than either of the full circles?

[NEW PAGE]

Now let's apply what we just learned. Suppose you are told that John Smith is a man 30 years of age who lives in the New Jersey area and that he once shoplifted when he was 16 years old. You also have the information that at that time he had many girlfriends. Consider the following questions: How likely is it that John Smith is a man with a criminal record? How likely is it that he is a father of two children? And finally, how likely is it that he is both a father of two children and has a criminal record? Most people would think that the probability that John Smith has a criminal record is quite high because they are given a piece of information about his past instance of shoplifting. For this reason, many people would assign a higher probability to the statement that John Smith is a father of two children and has a criminal record than to the statement that he is just a father of two children. Do you understand why this is wrong? To avoid responding in terms of what seems more probable instead of what is actually probable, you should think about how many more men in New Jersey are fathers of two children than are people with criminal records. The fact that John Smith had a shoplifting instance in his adolescence does not necessarily mean that he belongs to the group of people with criminal records. It would be helpful to you to consider the problem using the following illustration:

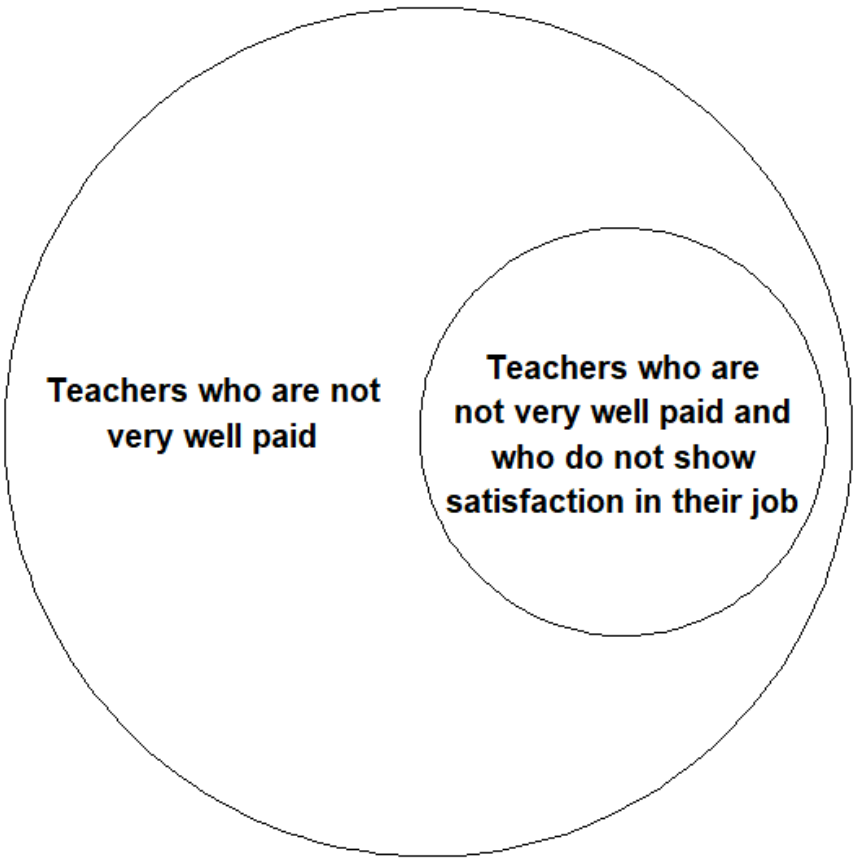


The diagram correctly assigns a larger circle to men who are fathers of two children and a smaller circle to men with criminal records. In addition, the shaded area indicates the portion of fathers of two children who also have criminal records. This shaded area is of a smaller size than either the full circle representing fathers of two children or the full circle representing people with criminal records.

[NEW PAGE]

As a final reminder, keep in mind that you can use these diagrams to evaluate the generality of a statement. For example, the statement “Most teachers are not very well paid” is more general than “Most teachers are not very well paid and do not show satisfaction in their jobs.” Because the first statement includes the second one, it is more general and, therefore, must

be more probable than (or at least equally probable as) the second one. In fact, when the second statement is true, the first one must be true, too. The use of a diagram representation in the way that has been suggested here would make the problem transparent. See how an appropriate diagram to represent the two statements would look:



Without Venn Diagrams

In this section, we are interested in studying how people reason and the kinds of difficulties they often encounter in solving problems. By studying the difficulties that people encounter in reasoning, we may be able to find general rules by which the human mind works.

Specifically, we are interested in studying how well people are able to learn how to reason about the frequencies of different categories or about the relationships between categories. For example, consider the relationship among three categories: (1) furniture, (2) chairs, and (3) kitchen chairs. You already know that "all kitchen chairs are chairs" and that "all chairs are a type of furniture."

As you can consider, the category of kitchen chairs must be smaller than the category of chairs, because we know that there are other types of chairs (for example, desk chairs and highchairs). Similarly, the category of chairs must be smaller than the category for furniture. There are, in fact, other types of furniture besides chairs (for example, tables, sofa, and desks).

[NEW PAGE]

Now, consider the relationship among the following three categories: (1) vehicles, (2) cars, and (3) Toyotas. Then, answer the following question.

Use your mouse to rank the following categories from smallest to largest:

___ Vehicles

___ Cars

___ Toyotas

Participant receives feedback on a new page: The vehicles category is larger than the cars category, which is larger than the Toyotas category.

[NEW PAGE]

The relationships that you have considered are for categories that are nested within each other. Sometimes, however, categories are completely non-overlapping. Think about these two categories: (1) U.S.-born women and (2) European-born women. Membership in one of these two categories excludes the possibility of belonging to the other category.

Another example is to consider the category of all women split into separate sections for women born in different parts of the world. These sections do not overlap. You may know that there are many more Asian-born than Australian-born women. At times, you may want to understand this fact by considering that larger groups will have higher frequencies and smaller groups will have lower frequencies. Other times, you don't have to think that hard about the sizes and you can just consider the inclusion or exclusion relations by the nesting or non-overlapping of the parts.

[NEW PAGE]

Now let's take one of these categories and consider how we can segment it. If we take the category of U.S.-born women, we can segment the entire category by age, creating, for example, eight subcategories of women who are 20 to 30, 30 to 40, 40 to 50, 50 to 60, 60 to 70, 70 to 80, 80 to 90, or 90 to 100 years of age.

Then, consider the subgroup of U.S.-born women, aged 20 to 30 and think of another subgroup nested within it. For example, we can divide this group by hair color. An appropriate subgroup would be all those U.S.-born women, aged 20 to 30 who have red hair. We can

continue this process by adding another characteristic: U.S.-born women, aged 20 to 30, red-haired, and weightlifters. As you may consider, the group of U.S.-born women who are 20 to 30 years of age and red-haired and weightlifters are the smallest category after all of this segmentation.

Sometimes, in order to make correct inferences it is important to accurately compare the size of two subgroups. If we ask you, for example, which of these two subgroups is larger,

Subgroup 1, U.S.-born women, 20 to 30 years of age
Subgroup 2, U.S.-born women, 20 to 30 years of age and red-haired,

you can use the logic discussed here to recognize that subgroup 1 must be larger than subgroup 2.

[NEW PAGE]

Let's try to practice what you have been learning. **Consider the following categories: (1) physicians and (2) lawyers.** For simplicity, let's say that these two categories are completely non-overlapping.

Then, in the category of physicians, consider how you would subdivide it into the two subcategories of married and unmarried physicians. Make a subsequent distinction within the subcategory of unmarried physicians by considering that some like to travel abroad, and some don't.

Now, of the following two subgroups, select what is larger:

___ Unmarried physicians
___ Unmarried physicians who like to travel abroad

[NEW PAGE]

Participant receives feedback: The subgroup of unmarried physicians is larger than the subgroup of unmarried physicians who like to travel abroad.

[NEW PAGE]

Finally, we want to point out to you that some categories can partially overlap without being identical. Consider the following two categories: (1) Professional women volleyball players and (2) Women who swim regularly for exercise.

The set of professional volleyball players who also swim regularly for exercise is a subset of both categories. There is a greater chance of being a woman who just swims for exercise than of being a woman who swims and is a professional volleyball player, too. Also, there is a greater chance of being just a professional women volleyball player than of being a volleyball player who also swims regularly for exercise. The overlapping subset of two categories must be smaller than either category.

[NEW PAGE]

Now let's apply what we just learned. Suppose you are told that John Smith is a man 30 years of age who lives in the New Jersey area and that he once shoplifted when he was 16 years old. You also have the information that at that time he had many girlfriends. Consider the following questions: How likely is it that John Smith is a man with a criminal record? How likely is it that he is a father of two children? And finally, how likely is it that he is both a father of two children and has a criminal record? Most people would think that the probability that John Smith has a criminal record is quite high because they are given a piece of information about his past instance of shoplifting. For this reason, many people would assign a higher probability to the statement that John Smith is a father of two children and has a criminal record than to the statement that he is just a father of two children. Do you understand why this is wrong? To avoid

responding in terms of what seems more probable instead of what is actually probable, you should think about how many more men in New Jersey are fathers of two children than are people with criminal records. The fact that John Smith had a shoplifting instance in his adolescence does not necessarily mean that he belongs to the group of people with criminal records.

Try considering the categories. The category of men who are fathers of two children is larger than the category of men with criminal records. In addition, the category of fathers of two children who also have criminal records is smaller than the category of men with criminal records. This category of fathers of two children who also have criminal records is smaller than either the category of fathers of two children or people with criminal records.

[NEW PAGE]

As a final reminder, keep in mind that you can use logic to evaluate the generality of a statement. For example, the statement “Most teachers are not very well paid” is more general than “Most teachers are not very well paid and do not show satisfaction in their jobs.” Because the first statement includes the second one, it is more general and, therefore, must be more probable than (or at least equally probable as) the second one. In fact, when the second statement is true, the first one must be true, too. The use of logic in the way that has been suggested here would make the problem transparent.

Appendix C: Supplementary Methods Section

URPP and Consent

There were 5 URPP participants who unregistered from the study and hence revoked their consent. This is also why there were 205 URPP participants who signed the consent form, despite there being only 200 participant-hours allocated for each experimenter.

Qualtrics and Spam

In Qualtrics, participants are marked as 'Spam', if the platform detects identical responses on the same survey from the same internet protocol (IP) address within 12 hours (Qualtrics Support, personal communication, March 29, 2022). However, it should be noted that my experiment could only detect spam based on responses in the consent form—for which identical responses across nearly all participants should be expected—and that I did not record IP addresses due to our ethics protocol. Only Prolific participants were marked as 'Spam' and their support team confirmed with me that the participants that were marked as 'Spam' did not have overlapping IP addresses with any of the other participants in my study (Prolific Support, personal communication, March 31, 2022 and May 5, 2022). Given this conflicting information from both sides, I elected to simply remove all participants marked as 'Spam'.

Power Analysis

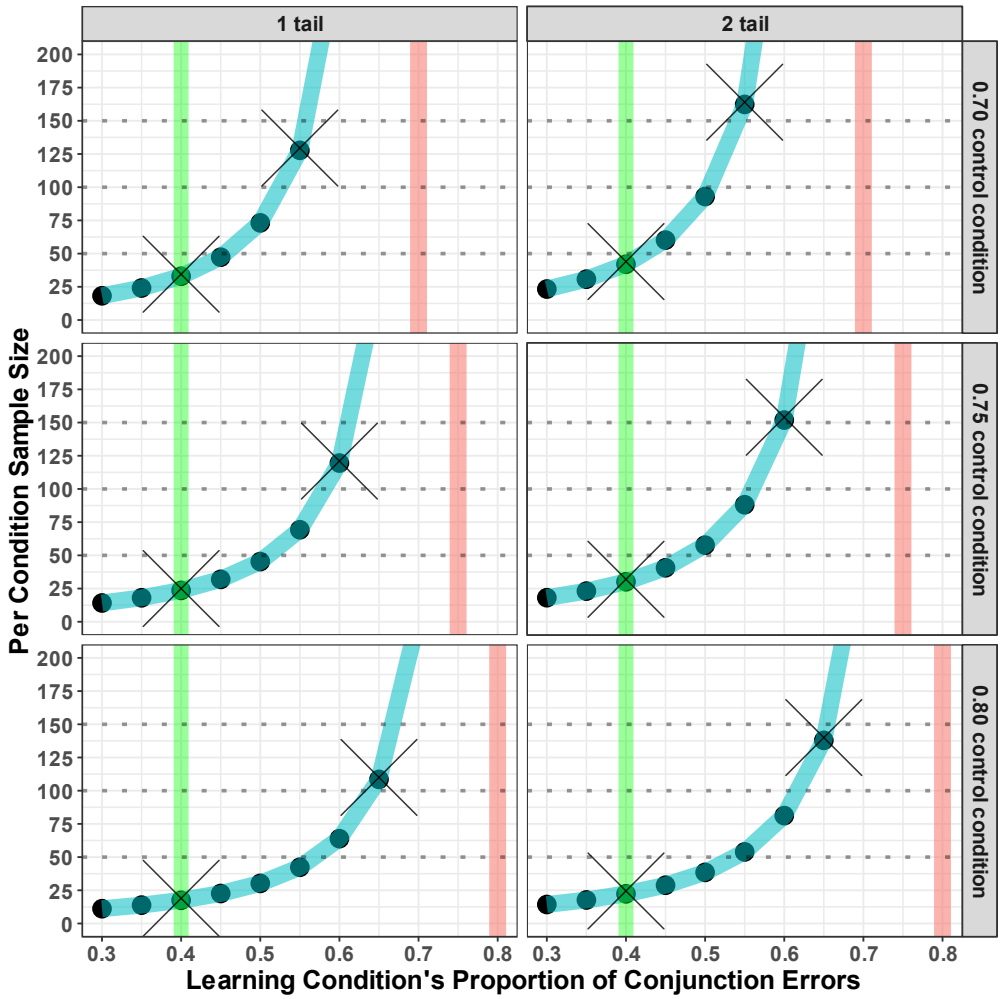
Before data were collected and as shown in Figure 12, I created a graph of a power analysis[15] using *G*Power* (Faul et al., 2007, 2009) to estimate the number of participants needed to detect a 15% difference in the proportion of conjunction errors on the Linda problem between the control and implicit training condition 80% of the time with a 5% chance of false positive

[15] Thank you to Dr. Robert Phil Chalmers for programming a Monte Carlo simulation of a closely related scenario to this problem.

error. I used a conservative baseline of between 70 to 80% of a sample, based on prior estimates of the proportion of conjunction errors without training (Hertwig & Chase, 1998; Tversky & Kahneman, 1983). I also depicted a liberal prior estimate of the proportion of conjunction errors in explicit training (Agnoli & Krantz, 1989; Fisk & Pidgeon, 1997) as a green vertical bar at 40% to get an estimate of the number of participants needed to replicate their study. Based on this graph and uncertainty in the number of participants we could test, I aimed to test between 60 and 165 participants per condition.

Figure 12

Pre-Experiment Power Analysis Graph of One- and Two-Sided z-Tests of Independent Proportions Facetted by Control Condition's Proportion of Conjunction Errors



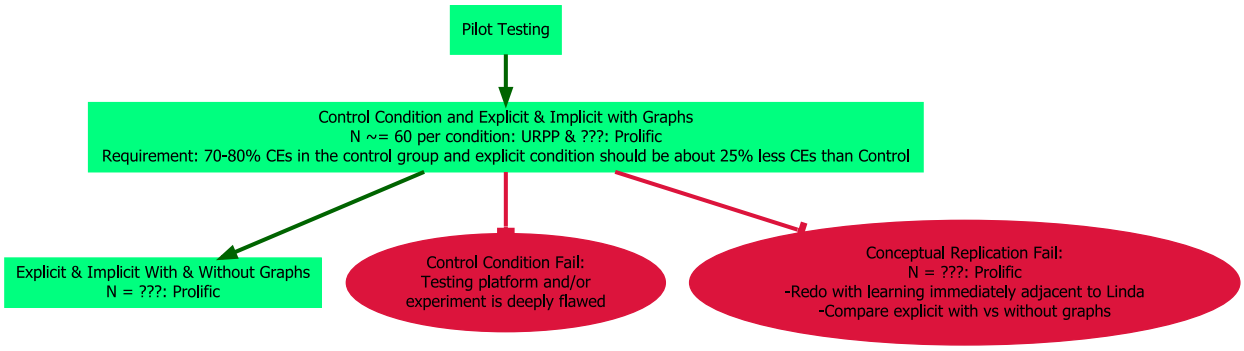
Note. Curve represents required sample size for 80% power and a type 1 error rate of 5%. Each row represents a different scenario based on the proportion of conjunction errors in the control condition. Red vertical line represents the control condition's proportion of conjunction errors and green, the explicit condition. Although logistic regression likelihood ratio tests were also used to test the main hypotheses because they have more power than Wald tests (Fox, 2016), I found it difficult to produce comparable power analysis graphs to the above for likelihood ratio tests using *G*Power*. Hence, the required sample size was compared with Wald tests from a binomial logistic regression—shown in large 'X' marks and as computed by *G*Power*—at the 'control minus-15%' and 40% proportion of conjunction errors to ensure some consistency between the *z*-tests visualized here and the logistic regression analyses that were ultimately used.

Stepwise Testing

Owing to uncertainty in how many participants' data could be collected, I also planned a decision flowchart to best balance the costs of the experiment and achieving its goals, as shown in Figure 13.

Figure 13

Pre-Registered Decision Flowchart for the Different Conditions Tested Based on Findings



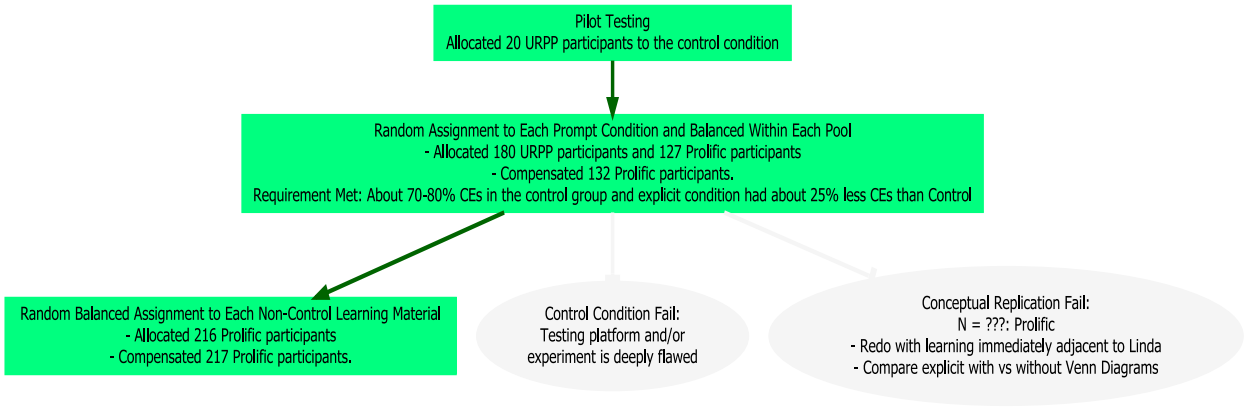
Note. Green arrows and boxes represent successful completions of requirements; red: failure to meet requirements.

Ultimately, there were three phases of data collection for this experiment as shown in Figure 13. Phase 1 represents the initial pilot testing in row 1 of Figure 13 – a quick survey of the results confirmed that the programmed experiment functioned as intentioned. Phase 2 is shown in row 2 of Figure 13. In phase 2, participants were randomly assigned such that the number of participants in each prompt condition would be balanced within each pool of participants (URPP or Prolific), meaning that there were twice as many control condition participants as in the sub-condition of an explicit prompt with graphs. Balancing across prompt conditions, as opposed to the crossing of the prompt and learning material conditions, was done to maximize the statistical power in testing objective 1. After confirming the conceptual replication and preliminary results

for objective 1, the remaining Prolific participants were randomly assigned across each learning material in phase 3 to maximize the statistical power for objectives 2 and 3.

Figure 14

Actual Stepwise Flowchart for Participant Allocations Based on Findings



Note. Because Prolific automatically substitutes participants who withdraw from the study (*Prolific*, 2022) and because all participants who sign the consent form were compensated, there were more compensated than allocated Prolific participants. Colored in green is the path this experiment took and in gray is the path not taken.

Mahalanobis Distance Variables

The following variables were used: *d*-prime scores; average log transformed reaction time on the VPA; median VPA reaction time; minimum VPA reaction time; median submission time on the CRT-7; mean reaction time on the CRT-7; minimum reaction time on the CRT-7; mean log-transformed reaction time on the CRT-7; submission time on the VNR; total time spent on either the reading comprehension or training task; median reaction time on the non-Linda problem parts of the heuristics and biases battery task; mean reaction time on the heuristics and biases battery task; mean log transformed reaction time on the heuristics and battery task; completion time on the Linda problem; and total experiment time.

Appendix D: Extra Results

Individual Differences

Table 10

Summary Statistics of CRT-7 Task by Prompt and Learning Material Level

Prompt	Learning Material	Baseline				Sensitivity			
		n	Mean (SD)	25%ile	75%ile	n	Mean (SD)	25%ile	75%ile
Control	Control	93	2.8 (2.33)	1	5	38	3.58 (2.15)	2	5.00
Implicit	Graphs & Text	96	2.96 (2.37)	1	5	68	2.93 (2.25)	1	4.25
Implicit	Text	88	3.16 (2.23)	1	5	63	3.51 (2.16)	1	5.00
Explicit	Graphs & Text	86	2.91 (2.45)	1	5	58	3.26 (2.51)	1	6.00
Explicit	Text	90	3.3 (2.31)	1	5	53	3.4 (2.27)	1	5.00

Table 11

Summary Statistics of VNR Task by Prompt and Learning Material Level

Prompt	Learning Material	Baseline				Sensitivity			
		n	Mean (SD)	25%ile	75%ile	n	Mean (SD)	25%ile	75%ile
Control	Control	93	5.49 (1.83)	4.00	7.00	38	6.08 (1.82)	5	7.00
Implicit	Graphs & Text	96	6.03 (2.13)	4.75	7.00	68	6.22 (2.19)	5	7.25
Implicit	Text	88	6.06 (1.94)	4.75	7.25	63	6.54 (1.78)	5	8.00
Explicit	Graphs & Text	86	5.76 (1.8)	5.00	7.00	58	6.02 (1.89)	5	7.00
Explicit	Text	90	5.83 (1.96)	4.00	7.00	53	6.38 (1.9)	5	8.00

Note. VNR scores are scored between 0 and 13. No participants scored 13 out of 13.

Table 12

Summary Statistics of VPA's d-Prime Measure by Prompt and Learning Material Level

Prompt	Learning Material	Baseline				Sensitivity			
		n	Mean (SD)	25%ile	75%ile	n	Mean (SD)	25%ile	75%ile
Control	Control	93	1.34 (1.06)	0.50	1.79	38	1.71 (1.02)	0.93	2.09
Implicit	Graphs & Text	96	1.22 (1.03)	0.48	1.84	68	1.38 (1.05)	0.53	2.10
Implicit	Text	88	1.15 (1.08)	0.32	1.79	63	1.13 (1.05)	0.32	1.79
Explicit	Graphs & Text	86	0.93 (0.98)	0.32	1.25	58	1.18 (0.98)	0.57	1.69
Explicit	Text	90	1.29 (1.31)	0.32	1.99	53	1.26 (1.3)	0.32	1.90

Table 13

Summary Statistics of VPA's Adjusted Hit Rate Measure by Prompt and Learning Material Level

Prompt	Learning Material	Baseline				Sensitivity			
		n	Mean (SD)	25%ile	75%ile	n	Mean (SD)	25%ile	75%ile
Control	Control	93	0.72 (0.16)	0.59	0.84	38	0.76 (0.15)	0.66	0.91
Implicit	Graphs & Text	96	0.71 (0.18)	0.59	0.84	68	0.73 (0.18)	0.59	0.84
Implicit	Text	88	0.72 (0.16)	0.59	0.84	63	0.72 (0.16)	0.59	0.84
Explicit	Graphs & Text	86	0.65 (0.19)	0.59	0.78	58	0.7 (0.18)	0.59	0.83
Explicit	Text	90	0.7 (0.21)	0.59	0.84	53	0.69 (0.22)	0.53	0.91

Reading Comprehension and Learning Task

Table 14

Summary Statistics of Reading Comprehension and Learning Material Responses

Prompt	Learning Material	Baseline				Sensitivity			
		n	Mean (SD)	25%ile	75%ile	n	Mean (SD)	25%ile	75%ile
Control	Control	93	5.02 (2.12)	4	7	38	6.71 (0.73)	6	7
Implicit	Graphs & Text	96	1.8 (0.43)	2	2	68	2 (0)	2	2
Implicit	Text	88	1.75 (0.49)	2	2	63	2 (0)	2	2
Explicit	Graphs & Text	86	1.72 (0.52)	2	2	58	2 (0)	2	2
Explicit	Text	90	1.74 (0.49)	2	2	53	2 (0)	2	2

Note. As shown in Figure 5, only control group participants complete the reading comprehension task. All other participants complete an adaptation of the learning materials provided by Agnoli and Krantz (1989), which has two multiple choice questions each scored as correct or incorrect.

Proportion of Errors in the Sensitvity Dataset

Table 15

Proportion of Each Condition That Made Conjunction Errors in the Sensitivity Dataset

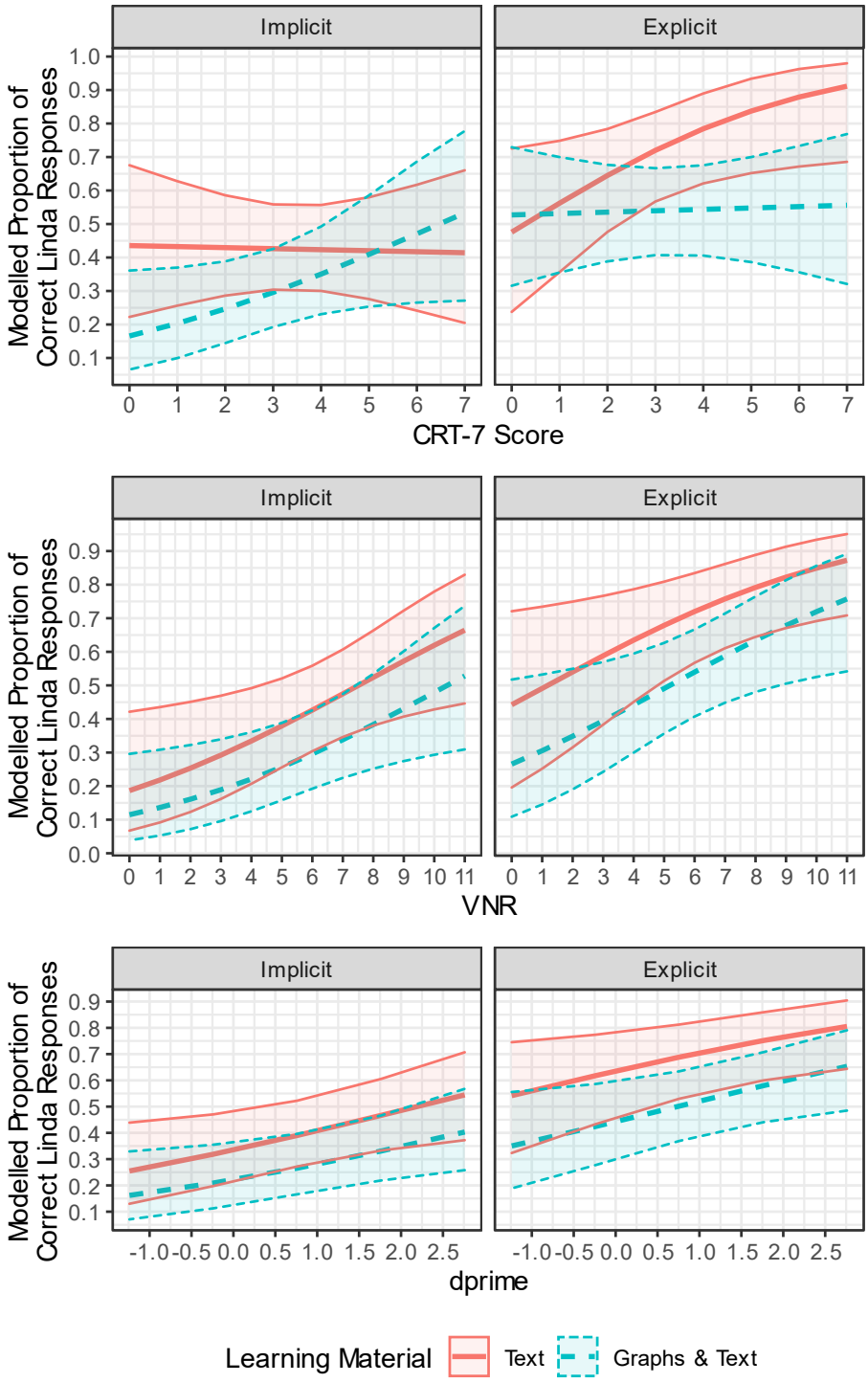
Prompt	Learning Material		
	Control	Graphs & Text	Text
Control	0.82 (0.06)		
Implicit		0.66 (0.06)	0.56 (0.06)
Explicit		0.47 (0.07)	0.28 (0.06)

Note. Computed in a similar manner as Table 4.

Objective 3 Sensitivity Regression Model

Figure 15

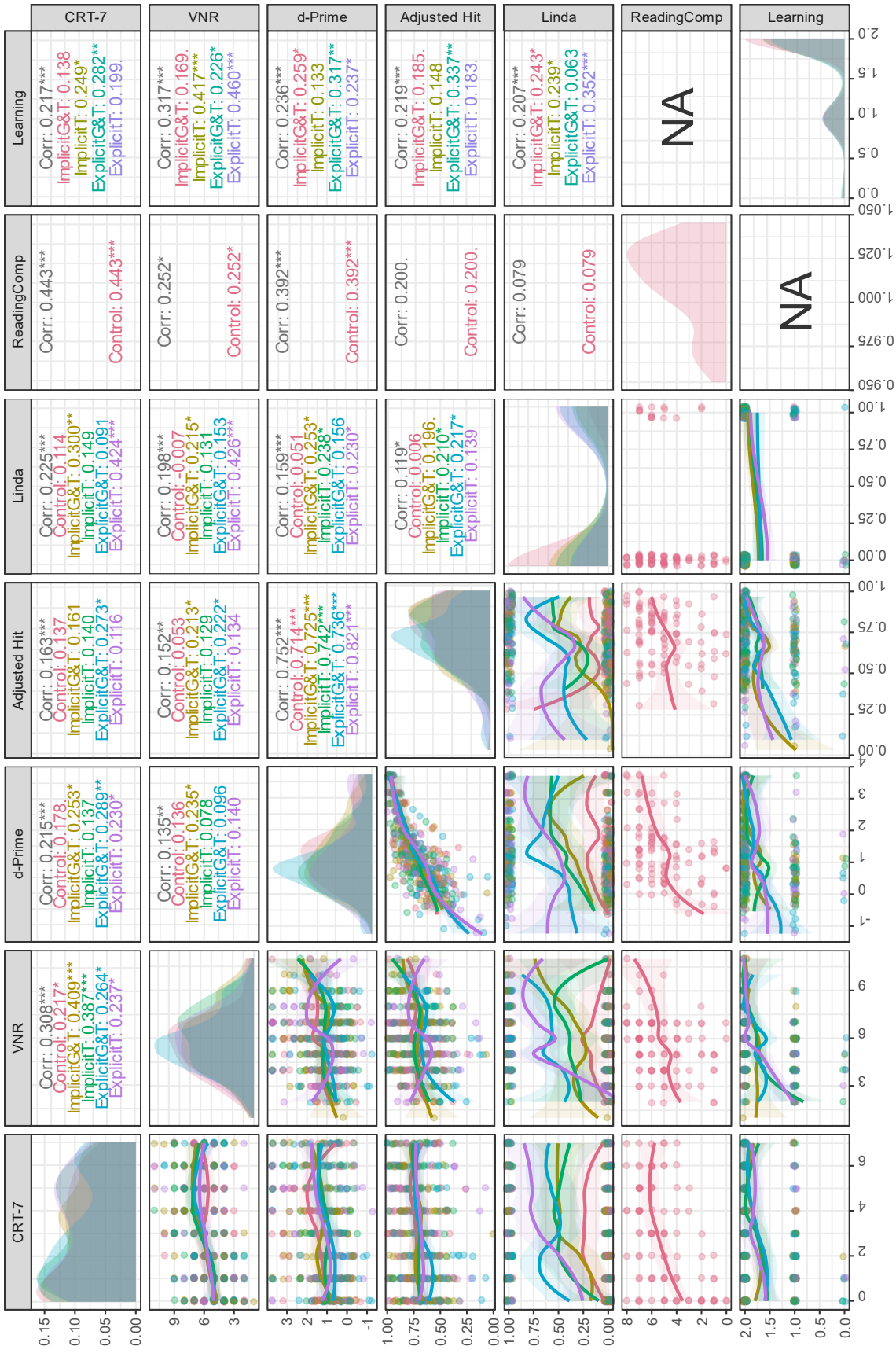
Visualization of Objective 3 Sensitivity Regression Model



Bivariate Intertask Correlation Matrix

Figure 16

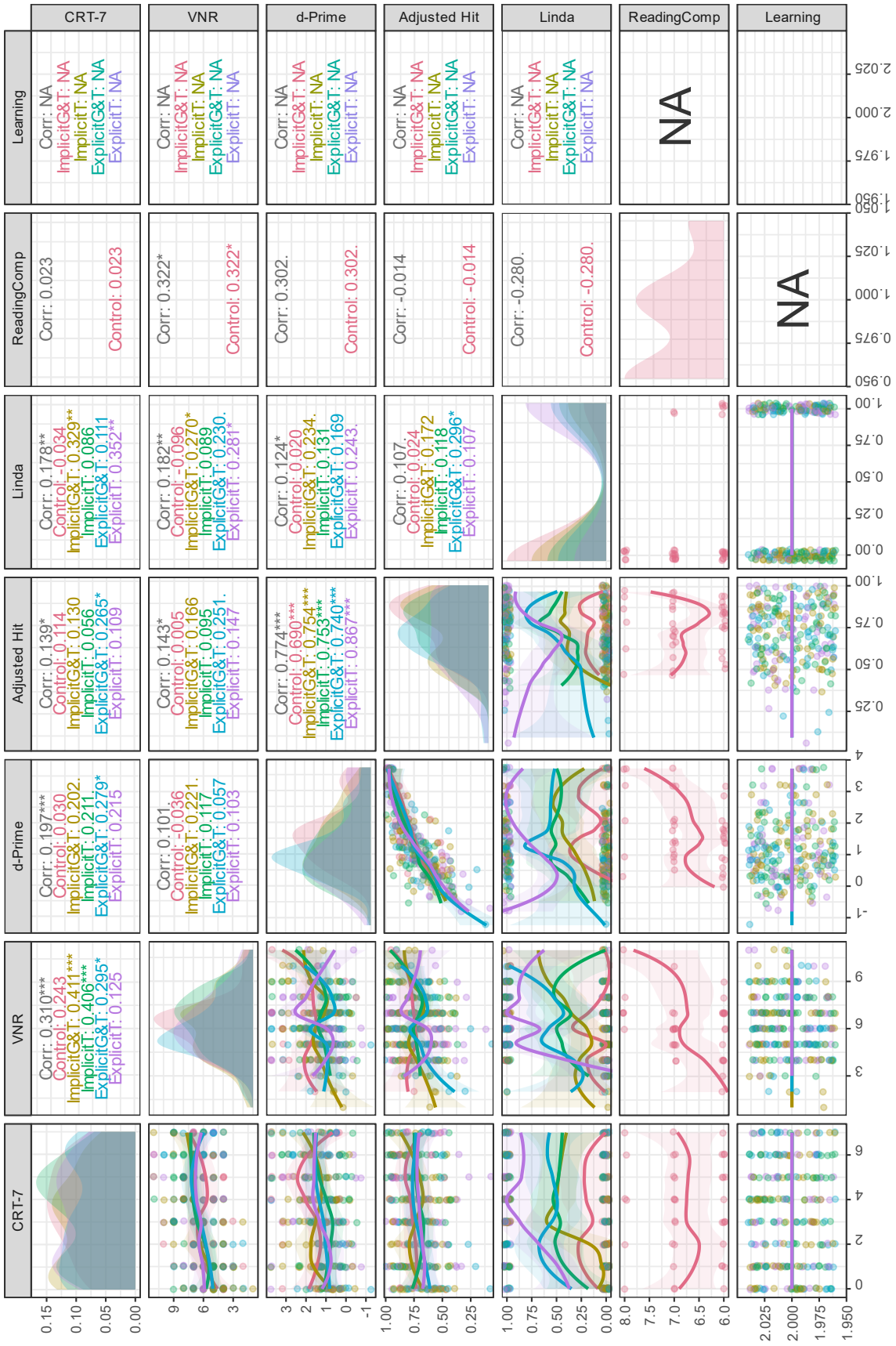
Correlation Matrix Using Baseline Data


CRT-7
VNR
d-Prime
Adjusted Hit
Linda
ReadingComp
Learning
Corr: 0.308***
Control: 0.217*
ImplicitG&T: 0.409***
ImplicitT: 0.387***
ExplicitG&T: 0.264*
ExplicitT: 0.237*
Corr: 0.215***
Control: 0.178
ImplicitG&T: 0.253*
ImplicitT: 0.137
ExplicitG&T: 0.289**
ExplicitT: 0.230*
Corr: 0.163***
Control: 0.137
ImplicitG&T: 0.161
ImplicitT: 0.140
ExplicitG&T: 0.273*
ExplicitT: 0.116
Corr: 0.225***
Control: 0.114
ImplicitG&T: 0.300**
ImplicitT: 0.149
ExplicitG&T: 0.091
ExplicitT: 0.424***
Corr: 0.443***
Control: 0.443***
Corr: 0.217***
ImplicitG&T: 0.138
ImplicitT: 0.249*
ExplicitG&T: 0.282**
ExplicitT: 0.199.
Corr: 0.135**
Control: 0.136
ImplicitG&T: 0.235*
ImplicitT: 0.078
ExplicitG&T: 0.096
ExplicitT: 0.140
Corr: 0.152**
Control: 0.053
ImplicitG&T: 0.213*
ImplicitT: 0.129
ExplicitG&T: 0.222*
ExplicitT: 0.134
Corr: 0.198***
Control: -0.007
ImplicitG&T: 0.215*
ImplicitT: 0.131
ExplicitG&T: 0.153
ExplicitT: 0.426***
Corr: 0.252*
Control: 0.252*
Corr: 0.317***
ImplicitG&T: 0.169.
ImplicitT: 0.417***
ExplicitG&T: 0.226*
ExplicitT: 0.460***
Corr: 0.752***
Control: 0.714***
ImplicitG&T: 0.725***
ImplicitT: 0.742***
ExplicitG&T: 0.736***
ExplicitT: 0.821***
Corr: 0.159***
Control: 0.051
ImplicitG&T: 0.253*
ImplicitT: 0.238*
ExplicitG&T: 0.156
ExplicitT: 0.230*
Corr: 0.392***
Control: 0.392***
Corr: 0.236***
ImplicitG&T: 0.259*
ImplicitT: 0.133
ExplicitG&T: 0.317**
ExplicitT: 0.237*
Corr: 0.119*
Control: 0.006
ImplicitG&T: 0.196.
ImplicitT: 0.210*
ExplicitG&T: 0.217*
ExplicitT: 0.139
Corr: 0.200.
Control: 0.200.
Corr: 0.219***
ImplicitG&T: 0.185.
ImplicitT: 0.148
ExplicitG&T: 0.337**
ExplicitT: 0.183.
Corr: 0.079
Control: 0.079
Corr: 0.207***
ImplicitG&T: 0.243*
ImplicitT: 0.239*
ExplicitG&T: 0.063
ExplicitT: 0.352***
NA
NA

Note. Each column and row lists the task that is being graphed. The intersection of two of the same task is shown in the diagonal as a densitiy plot. Plots in the lower left triangle are colored by learning material by prompt condition and show loess lines when possible. Statistics in the upper right triangle represent pearson correlations of the corresponding tasks along with asterisk for significance—the gray 'Corr' stands for the overall correlation. *** stands for $p < 0.001$, ** stands for $p < 0.01$, * stands for $p < 0.05$, . Stands for $p < 0.10$.

Figure 17

Correlation Matrix Using Sensitivity Data


CRT-7
VNR
d-Prime
Adjusted Hit
Linda
ReadingComp
Learning
Corr: NA
ImplicitG&T: NA
ImplicitT: NA
ExplicitG&T: NA
ExplicitT: NA
Corr: 0.023
Control: 0.023
Corr: 0.322*
Control: 0.322*
Corr: 0.302.
Control: 0.302.
Corr: -0.014
Control: -0.014
Corr: -0.280.
Control: -0.280.
NA