Title of Thesis:
Semantic Context Produces Overconfidence in One's Ability to Perform Highly Specialised Skills

A thesis
submitted in partial fulfilment
of the requirements for the degree
of
Doctor of Philosophy in Psychology
at
The University of Waikato
by
Kayla Jordan

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2022

Abstract

Some research suggests people are overconfident because of their personality characteristics, a lack of insight, or because overconfidence is beneficial in its own right (Dunning et al., 2003; Johnson & Fowler, 2011; Paulhus et al., 2003). But other research fits with the possibility that fluent experience in the moment can rapidly drive overconfidence (Alter and Oppenheimer, 2008). For example, fluency can push people to become overconfident in their ability to throw a dart, know how rainbows form or predict the future value of a commodity (Cardwell, Lindsay & Garry, 2017; Kardas and O'Brien, 2018; Newman et al., 2018). But surely there are limits to overconfidence. That is, even in the face of fluency manipulations known to increase feelings of confidence, reasonable people would reject the thought that they, for example, might be able to land a plane in an emergency or understand a foreign language. Across 11 experiments and 3529 subjects, we[1] provide evidence supporting the idea that semantic context[2] can rapidly increase people's confidence in their ability to perform these highly specialised skills.

In Part 1, we asked some people (but not others) to watch a trivially informative video of a pilot landing a plane before they rated their confidence in their own ability to land a plane. We found watching the video inflated people's confidence that they could land a plane. Moreover, we found people who watched the video were more confident in their ability to land the plane when they found the scenario easier to imagine. These findings fit with the idea that the semantic context of the video enables subjects to develop more detailed

[1] The research conducted in this thesis is my own. However, I conducted it in a lab and supervised a team of Honours students, and undergraduate students. I also received advice and direction from my supervisors. Therefore, I use the word "we" in this thesis to reflect those facts. Elsewhere, I use the word "we" to refer to what is known—or not known—in the wider scientific community.

[2] In this thesis—drawing on work from Bransford and Johnson (1972), Cardwell, Newman, et al. (2017), and Newman et al. (2012)—we define semantic context as: additional information a person has on hand about a particular concept or topic that helps them to understand and interpret new information related to that concept or topic.

imaginations of themselves landing the plane, that they misconstrue as evidence that they could actually land the plane themselves.

In Part 2, we found semantic context produced a similar illusion of ability when people were asked to evaluate their ability to apply foreign language abilities in new situations. We showed subjects a video clip of people speaking Danish either with, or without English subtitles. Then we asked subjects to rate how well they thought they would be able to understand Danish in new situations. Our results support the idea that semantic context boosts people's confidence by making it easier to understand the situation at hand, but people then mistake that ease as evidence of their ability to comprehend Danish in novel situations. Our findings extend prior work by suggesting that increased semantic context creates illusions not just of prior experience or understanding—but also of the ability to actually do something implausible (Whittlesea, 1993; Cardwell, Lindsay & Garry, 2017)[3].

[3] Portions of this thesis were adapted from:

Jordan, K., Zajac, R., Bernstein, D., Joshi, C., & Garry, M. (2022). Trivially informative semantic context inflates people's confidence they can perform a highly complex skill. *Royal Society Open Science*, *9*(3), 211977. https://doi.org/10.1098/rsos.211977

Jordan, K., Nielsen, N. P., Bernstein, D., Newman, E. J., Zajac, R. & Garry, M. (manuscript in preparation). Den mørke side af semantisk kontekst [The dark side of semantic context]: semantic context boosts people's confidence in their ability to comprehend Danish.

But I have expanded on the introduction, method, results, and discussion sections.

Acknowledgements

I'm extremely grateful for the funding from the University of Waikato for supporting this research. I'd also like to thank all of the subjects who participated in this research, and all of the undergraduate and honours students who helped with data collection, materials design, and provided helpful ideas about next steps.

To the Garry Labsters: I feel so lucky to have such a supportive and encouraging academic family. Thank you all for your thoughtful feedback, troubleshooting sessions, and for teaching me all of the things. A special shoutout goes to Andrea Taylor for being such an incredible mentor, role model, and friend.

To Dan Bernstein, Eryn Newman, Niels Peter Nielsen, and Chaitanya Joshi: I couldn't have asked for more wonderful people to do this research with. Thank you for all your brilliant insights and feedback. This work wouldn't have been possible without your contributions.

To Rachel Zajac: it's been such a pleasure having you on my supervisory panel. Thank you for all your encouragement, for helping me transform clunky sentences and difficult transitions into beautiful prose, and for always getting my drafts back to me in record time.

To Maryanne: there aren't enough words to express how grateful I am to have had such an encouraging, enthusiastic and intelligent supervisor. Thank you for pushing me out of my comfort zone, introducing me to so many incredible people, providing advice and endless opportunities, spurring my love for research, and constantly reminding me about the importance of apostrophes. I'll be forever grateful that I didn't let my trip to Hawaii stop me from taking your directed study course 5 years ago.

To my family and friends, thank you for all your love, support, and encouragement. I'm blessed to have such a wonderful support system. To Logan: thank you for being my best friend and for supporting me through every stage of this process.

Table of Contents

List of Figures: Part 1

List of Tables: Part 1

List of Figures: Part 2

List of Tables: Part 2

Part 1

Chapter 1

Imagine this: you're on a small commuter plane when the pilot becomes incapacitated. You are the only person left to land the plane. How confident are you that you could land the plane without dying? It seems obvious that for anyone without flight experience, confidence should be vanishingly small. After all, landing a plane is a highly specialized skill, requiring hundreds of hours of training and a deep understanding of physics, engineering, and meteorology (Federal Aviation Administration, 2016). But here we present evidence that a disturbing proportion of ordinary people display some confidence they could land a plane safely and become even more confident after we show them one trivially-relevant piece of information: a 3-minute video of a pilot landing a plane. How can people come to be so overconfident in their skills, even highly specialised skills that require years of training? The psychological literature provides two explanations that partly—but not fully—help us to understand how people come to be overconfident in their abilities. In this thesis, we address a third explanation that has surprisingly remained unaddressed.

Existing explanations for overconfidence

Overconfidence depends on characteristics of the person

The first of these two explanations ties overconfidence to characteristics of the overconfident person. In the personality literature, for example, the degree to which people are overconfident depends on relatively stable features, such as personality traits. More specifically, one type of overconfidence called “overclaiming”—or claiming to know terms that do not exist—can be a form of social desirability, self-enhancement, or even narcissism (Paulhus et al., 2003: Paulhus & Harms, 2004). Overclaiming is typically measured by presenting subjects with a list of terms and asking them to rate their familiarity with each

term. Three of every 15 of these terms are foils—they don't exist. For example, the foil "plates of parallax" would be embedded within a list of real terms such as "satellite" and "asteroid". Therefore, anyone who indicates familiarity with a foil is overclaiming. It is common for people to falsely recognise several of these foils, even though they have never seen them before. What's more, the extent to which people overclaim is used to detect forms of self-enhancement in personality questionnaires (Paulhus, 2011; Paulhus et al., 2003). In other words, people "overclaim" what they know, or what they can do, so other people think better of them. The literature, then, supports the idea that personality traits such as narcissism play some role in a person's tendency towards overconfidence.

But it is not only personality traits that are associated with overconfidence. We also know from the social psychology literature that people's overall level of skill plays a role in their susceptibility to overconfidence. For example, we know that across a wide range of skills—including humour, grammar, and logic—that the worst performers overestimate their abilities the most (Dunning, 2011; Kruger & Dunning, 1999). The proposed mechanism driving this effect, the well-known "Dunning Kruger effect," is a metacognitive failure whereby the least skilled people are "ignorant of their own ignorance" (Dunning et al., 2003; Hodges, Regehr, & Martin, 2001). Not only can they not execute the target skills effectively, but they also lack the ability to evaluate their own level of skill accurately (Dunning et al., 2003; Simons, 2013).

Taken together, then, both the personality and social psychology literature suggest that when people are overconfident about what they know or what they can do, at least some of the time that overconfidence is driven by relatively stable characteristics, such as their personality, or their overall level of skill. But people do not have to be extremely narcissistic or the worst performers to be overconfident—in fact, most people are prone to these miscalibrated judgements about their knowledge or ability, particularly when comparing

themselves to the "average" person (Alicke & Govorun, 2005). These "above average" effects have been demonstrated across a diverse range of domains (Cross, 1977; Williams, 2003; Zenger, 1992). For example, 90% of drivers rate themselves as better than the average driver; 90% of teachers think they above average; and 32-42% of software engineers think that they are among the top 5% of employees in their companies (Cross, 1977; Williams, 2003; Zenger, 1992). It is impossible that all of these people are correct. This "above average" bias is thought to be driven by people's miscalibrated estimates of their own ability rather than a miscalibrated assessment of the abilities of others (Engeler & Häubl, 2021).

Finally, overconfidence is associated with gender (Barber & Odean, 2001, Beyer, 1990). Men tend to be more overconfident in their knowledge and abilities than women—even in a high-stakes environment, such as competitive running and diving (Krawczyk & Wilamowski, 2017; Lackner & Sonnabend, 2020). This gender overconfidence gap is most prevalent when people are asked to evaluate their performance on a masculine-gender-typed task—though women do not show the same overconfidence for feminine-gender-typed tasks (Beyer, 1990).

These explanations suggest overconfidence arises as a byproduct of, or associated with, personality traits or lack of insight. But a second literature suggests that overconfidence is beneficial in its own right.

Overconfidence is beneficial

We know from the literature on social cognition that believing that we are above average or holding unrealistically positive evaluations about ourselves is a key part of the "positive illusions" most of us have (Taylor & Brown, 1988). For example, when people are overconfident in their abilities to do something, the overconfidence can bias them to focus on the expected benefits rather than the costs (Johnson & Fowler, 2011; McKay & Dennet, 2009). This focus on benefits may in turn drive people to pursue

more difficult goals, to put in more effort and persistence to accomplish those goals, and sometimes even to accomplish them (Johnson & Fowler, 2011; Taylor & Brown, 1988). In fact, overconfident CEOs who invest in risky projects have also been found to be better innovators (Hirshleifer et al., 2012). There are also social benefits to overconfidence. For instance, people think that overconfident others are more knowledgeable and more trustworthy than their equally competent—but not overconfident—peers, and award them higher status (Anderson et al., 2012; Johnson, Wrangham & Rosen, 2002; Sniezek & Van Swol, 2001). Considered together, these literatures suggest that a little self-enhancement does not hurt in most cases; in fact, there are often numerous benefits (Johnson & Fowler, 2011).

If we step back now, we see that in some cases, overconfidence can be beneficial, and the extent to which people are overconfident can be influenced by relatively stable characteristics of a person, such as personality and overall level of skill. But the explanations for overconfidence covered this far are insufficient because they ignore the context of a situation—what is going on right now in this moment.

Feelings as information

There is reason to think that experience in the moment can rapidly drive overconfidence in one's knowledge about the world, performance, or even their future. For example, we know that people's beliefs and expectations about themselves, others, and the world are often influenced by cognitive biases in the moment (Cardwell, Lindsay & Garry, 2017; Newman et al., 2012; Schwarz, 2004). More specifically, when we process information, we are influenced by how easy or difficult it feels to do so. We often interpret easy processing, or *fluency*, along positive dimensions, such as truth, attractiveness, or understanding (Jacoby & Dallas, 1981; Oppenheimer, 2008). For example, when people are asked how well they understand how a specific—often

complex— process works (such as how rainbows form), people tend to report a better understanding of how that process works if the question is accompanied by an uninformative photo (say, a rainbow)—even though subsequent analysis shows no such improved understanding (Cardwell, Lindsay & Garry, 2017). These photos might provide a semantic context that increases fluency and makes it easier for people to bring related information to mind. That is, although the photo provided no useful information to inform their judgement, people might attribute some meaning to it due to ease of processing, and in turn interpret this feeling of ease as evidence of understanding (Schwarz, 2018; Unkelbach & Greifeneder, 2013; Wilson & Westerman, 2018).

Related work shows that simply altering the way questions are presented can change people's evaluations of their performance in the moment. In one line of experimentation, people thought they performed better on a test when the items were ordered from easiest to hardest, rather than the reverse—but in reality, performance was the same (Weinstein & Roediger, 2010, 2012). In a follow-up study, people evaluated their confidence in each answer as they worked through items on a test—and then, retrospectively, estimated their performance after each block of 10. The order of items did not change subjects' item-by-item confidence ratings, but it did change subject's estimated performance on the blocks. That is, "easy to hard" subjects rapidly, and retrospectively, became more optimistic about their performance on blocks of items compared with "hard to easy" subjects. Evidence suggests these patterns occur because people generate initial beliefs about their performance based on how easy or difficult it feels to answer the first few test questions—then, as the questions become more (or less) difficult, people do not sufficiently adjust those beliefs away from their initial anchor (Michael & Garry, 2019; Tversky & Kahneman, 1974; Weinstein & Roediger, 2012). Considered together, these findings suggest that situational factors, such as the order of items, that change how easy or difficult it feels to process information can

change people's evaluations of their performance on a task even though these evaluations do not reflect reality.

But fluency does not only influence people's judgements about knowledge. Instead, fluency can influence a diverse range of judgements in predictable ways—including judgements of truth, confidence, attractiveness, liking, and the difficulty of a task (Hertzog et al., 2003; Kardas & O'Brien, 2018; Kelley & Lindsay, 1993; Schwartz & Metcalfe, 1992). For example, people think that certain recipes are easier to follow and exercise routines easier to do when they are presented in an easy to read font, like Arial, than a difficult to read font, like Brush (Song & Schwarz, 2008). In another study, people who watched demonstrations of magic tricks or dart-throwing 20 times become more confident they could do those same tasks compared with people who watched the demonstrations only once (Kardas & O'Brien, 2018). A diverse range of manipulations have similar effects on people's judgements (see Alter & Oppenheimer, 2008 for a review). Let's consider conceptual manipulations on people's judgements of truth, for example: people are more likely to say a statement is true when it rhymes, when it is paired with a photo, and when words or concepts related to that statement are semantically primed (Begg, Anas, & Farinacci, 1992; Kelley & Lindsay, 1993; McGlone & Tofighbakhsh, 2000; Reber & Schwarz, 1999). Although these manipulations do not give people any information about whether the statement is true, they all help to enrich the context, a situation that makes it feel easier for people to bring to mind related information (Unkelbach & Greifeneder, 2013). Conversely, it is possible to manipulate fluency by changing how easy it is for people to physically perceive information. Manipulating perceptual features—like the clarity of visual stimuli by presenting those stimuli in high contrast, or auditory stimuli by changing the clarity of the speakers voice—produce processing ease, which people misattribute as evidence of truth or credibility (Reber & Schwarz, 1999, Newman & Schwarz, 2018).

Why do people interpret these feelings of ease along positive dimensions such as truth, attractiveness or understanding? One possibility is a two-stage process. First, over the course of our lives, we learn to associate easy processing with "things" that are familiar; second, we come to associate familiarity with various positive attributes (Unkelbach, 2006). There are two related explanations, in two different literatures, for why that happens. The cognitive literature suggests that we associate familiarity with (for example) truth instead of falsity, because we are exposed many more times to true statements than to any one of the corresponding false statements (Unkelbach, 2007). Therefore, Unkelbach suggests, people learn that familiarity is a signal of truth. The social literature suggests that we are predisposed to associate feelings of ease and familiarity with positive attributes because in our evolutionary past, it was beneficial to be cautious of novelty (Winkielman et al., 2003). When our ancestors experienced feelings of familiarity, suggest Winkielman and colleagues, this feeling was an important indicator of safety, because it would have meant that the last time they encountered that "thing," they survived that encounter. Therefore, we have evolved to adopt a heuristic, or mental shortcut, that familiar "things" are positive.

This fast, effortless heuristic processing is one of the two types of processing people use when making decisions and determining the source of their thoughts and feelings (Lindsay, 2008). And because heuristic processing is fast and effortless, it is often the default processing (versus slow, deliberate, and effortful "strategic" processing; see also Chaiken, 1980). As a result, people might use heuristics, such as fluency, even when they should not (Haselton, Nettle & Murray, 2015). Therefore, by manipulating the feeling of how easy it is to process some "thing" we can encourage people to find that "thing" true, attractive, understandable, or real—even when such a conclusion would be wrong.

Taken together, research suggests that fluency consistently influences people's judgements in the moment, in various ways. It is most influential when the situation

encourages heuristic processing, such as when it feels easier to bring to mind information related to the judgement at hand. Therefore, if we can manipulate a situation so that people can easily bring to mind information about performing a skill, they should quickly become overconfident in their ability to perform that skill.

Source monitoring

There is another way to consider the possibility that people would be more overconfident in situations where information is easily brought to mind: through the lens of the Source Monitoring Framework (SMF; Johnson et al, 1993). Whenever people bring information to mind, they make a seemingly automatic decision about where that information came from; to do this, they rely on characteristics such as ease of imagery and detail to determine the source (Johnson et al, 1993; Lindsay, 2008). Most of the time this source monitoring process works well, but people are more prone to make errors when the characteristics of one source overlap with the characteristics of another source (Lindsay, 2014). For example, when people see trivia statements with accompanying pictures, those pictures make it easier for people to bring to mind related thoughts, feelings, and images, thereby increasing characteristics such as concreteness, vividness and detail, that people construe as evidence that the claim is true (Newman 2018). Put another way, source monitoring errors such as these arise when people attribute thoughts or feelings from one source (here, ease of bringing to mind thoughts, vivid images, and feelings related to the picture) to another source (truth or memory; having seen the statement before). The SMF also provides a rationale for people making similar errors in evaluations about the future—findings we elaborate on below (for a review of the SMF, see Lindsay, 2014). A classic paper suggests that both the past and future "are constructions of the present," with the past constantly changing and the future constrained by what we are doing right now (Johnson & Sherman, 1990 p.483). These ideas fit with modern evidence suggesting that imagination and

memory are “fundamentally the same process” (Addis, 2020 p.233). Indeed, we know people rely on similar characteristics when judging the past and when estimating the likelihood of some event in the future (Atance & O'Neill, 2001; Johnson et al., 1993; Schacter & Addis., 2007; Suddendorf & Corballis, 2007; Tulving, 1985). Just like their past counterparts, future events are thought more likely when those events come to mind easily, are full of sensory detail, and are low on cognitive markers of effort (see Johnson & Sherman, 1990, and Schacter et al., 2007, for related views).

Boundary conditions to overconfidence

But surely there are limits to overconfidence. There is a vast difference, for example, between landing a plane and assessing confidence in one’s ability to cook a dish, throw a dart, or know how rainbows form. When it comes to appraising highly-specialized skills, such as landing a plane, perhaps people can take more reliable shortcuts than turning to feelings as information. Landing a plane is widely known to require expertise that arises from dedicated training and practice. It seems reasonable to speculate that—even in the face of fluency manipulations known to increase feelings of confidence—people might find this fact easily accessible, resulting in more realistic estimates of ability.

Nonetheless, some real-world survey data suggest that people, especially men, can be remarkably overconfident in their ability to do ridiculous things. Data from a recent YouGov survey showed 12% of men claimed they could win a point in a game against 23-time tennis grand slam winner Serena Williams (Yougov, 2019). Only 3% of women made this claim. And lest we think this disparity is a single cherry-picked example from the competitive world of sports, in another YouGov survey asking men and women to identify which animals they could beat in a fight, more men than women claimed they could beat every animal. This list included King Cobras, bears, and eagles

(Yougov, 2021). In short, these preposterous examples suggest overconfidence might emerge as well when people are asked if they could land a plane in an emergency. Such a prediction is supported more broadly by a growing experimental literature that suggests that even when people have demonstrated their knowledge of something in the past (say, by providing the correct answer to a trivia question), they don't always apply that knowledge in the future (Marsh & Umanath, 2014). In a classic example, people first take a test of general knowledge, and then, in a follow-up session, take a new test. This time, they are told to answer the questions only when the stem contains correct information, such as "Who found the glass slipper left at the ball by Cinderella?" (Erickson and Mattson, 1981; Reder & Kusbit, 1991). But many people fail to detect errors and mistakenly answer questions with an incorrect stem, such as "How many animals of each kind did Moses take on the ark?"—even when their answers on the first test showed they knew Noah was the correct biblical reference (Erickson and Mattson, 1981). These findings suggest that even in the face of genuine awareness that landing a plane requires expertise, people might still come to ignore that awareness in certain situations.

Considered together, several literatures converge on the possibility that, if we increase semantic context by showing people a related but uninformative video of a pilot landing a plane, they could come to rapidly display increased confidence in their ability to land a plane safely. These findings would suggest significant boundary conditions on prior demonstrations of overconfidence.

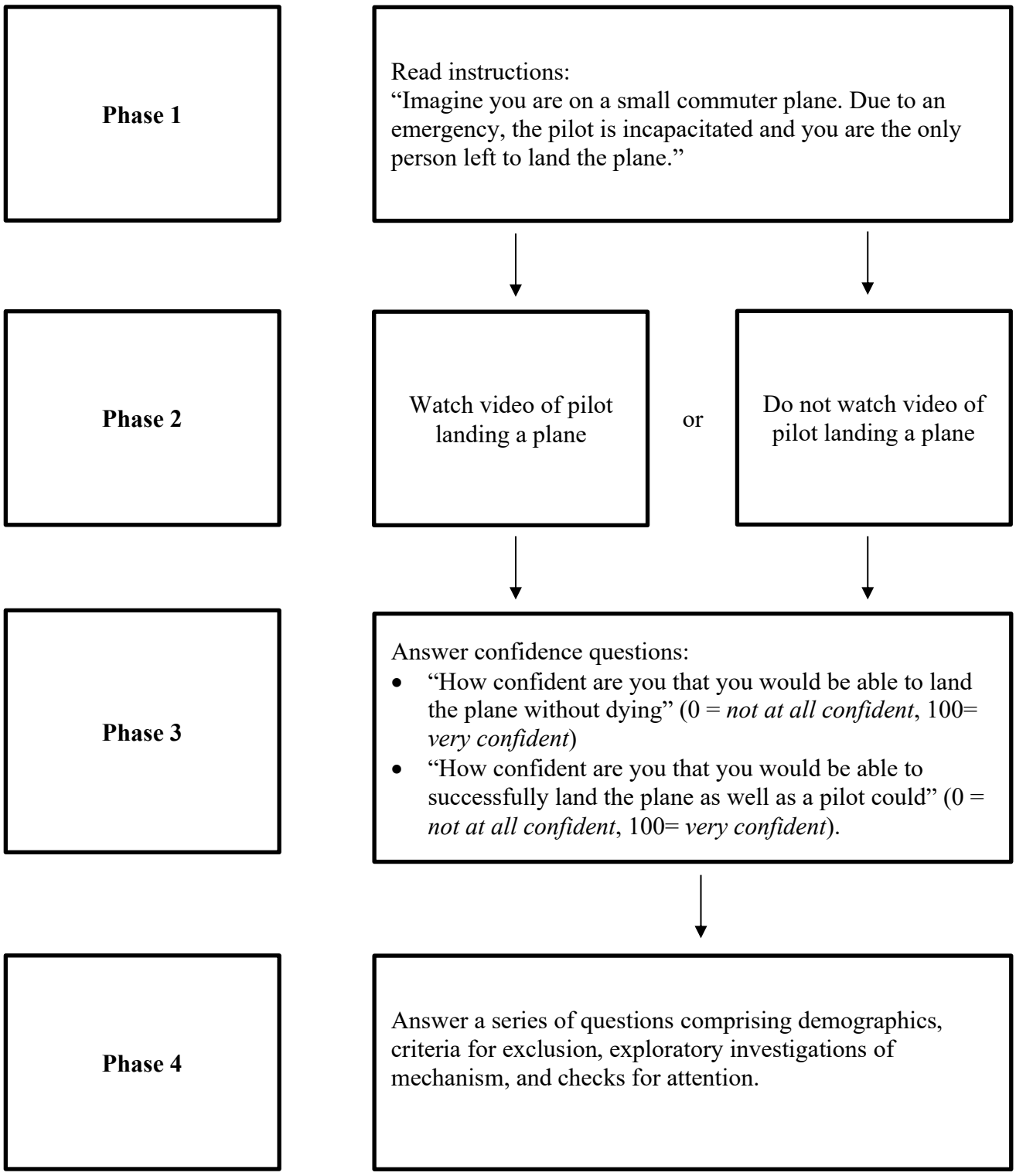
Overview of experiments

To what extent does watching a video of an expert performing a skill inflate people's confidence in their own ability to perform that skill? To answer this question, in Experiment 1, we asked some people to watch a short, non-instructional video of a

pilot landing a plane. Others did not watch the video. Immediately afterwards, everyone rated their confidence in their own ability to land a plane to a high standard (as well as a pilot could) and to a lower standard (without dying). In Experiment 2, we addressed the possibility that the order in which these two questions were presented mattered. In Experiments 3a and 3b, we provide support for one potential mechanism for producing this illusion: that watching videos increases people's ease of imagining themselves landing a plane.

Figure 1

Overview of the general method. The boxes summarise the key components across four phases in Experiment 1 and Experiment 2.



Chapter 2

Experiment 1

Method

Subjects. We recruited subjects from Amazon's Mechanical Turk (MTurk), an online source of a vast and diverse population, who received $0.30 USD for participating (see Mason & Suri, 2012). In the absence of data to estimate the true size of an effect, we instead aimed to collect data from 100 subjects in each condition. Anticipating exclusions, we collected data from 225, and retained a total of 198 subjects after exclusions (M_{age} = 40.47, SD_{age} = 13.42; 37% identified as men, 62% identified as women, and 1% identified as gender diverse); 98% of subjects reported English was their first language.

Design. We used a between-subjects design with two conditions (Video: video, no video).

Procedure. The experiment proceeded in four phases. First, subjects read the following instruction: "Imagine you are on a small commuter plane. Due to an emergency, the pilot is incapacitated and you are the only person left to land the plane." Second, subjects were randomly assigned to watch either a video of a pilot landing a plane, or to watch no such video. This video was 3m 44s and had no audio track. It was filmed from the back of the flight deck, and depicted the view of the flight deck, and the back of the pilots. Although it was clear the pilots were engaging with the instruments and controls, there were only obstructed views of their hands and what they were doing. In short, the video did not teach people how to land a plane, and had little instructional merit; moreover, when we showed the video to a retired airline captain who flew for over 35 years with Air New Zealand and taught other pilots, he pronounced the video "100% useless." In line with prior work on "truthiness," we

therefore classified the video as "non-probative"—it was related to the task at hand, but did not teach anyone anything about how to do the task (Newman et al., 2012; Cardwell, Lindsay & Garry, 2017).

In the third phase, subjects were told:

> Now we're going to ask you a few questions. Don't try to analyze and puzzle things out—just go with your gut feel or hunch. Respond as quickly as possible within a couple of seconds. Remember this is an emergency situation.

Then everyone answered the following confidence questions, in a fixed order because for our first attempt to answer our research question, we aimed to maximize the size of effects by presenting the "low bar" question first (Schwarz, 1999; Tversky & Kahneman, 1974): "How confident are you that you would be able to land the plane without dying?" (0 = *Not at all confident*, 100 = *Very confident*) and "How confident are you that you would be able to successfully land the plane as well as a pilot could?" (0 = *Not at all confident*, 100 = *Very confident*). We then asked subjects a series of questions comprising criteria for exclusion ("Are you a pilot?" [Yes, No], "Have you flown a plane before?" [Yes, No], and "Have you ever landed a plane before?" [Yes, No]), exploratory investigations of mechanism, ("How difficult was it for you to imagine attempting to land the plane?" [1 = *Not at all difficult*, 5 = *Very difficult*]; and "How much expertise do you think is involved in landing a plane?" [1 = *No expertise*, 5 = *A great deal of expertise*]); and checks for attention ("Which of the following best describes the situation you were asked to imagine?" [Performing emergency surgery, Going to see a doctor, Taking a flying lesson, Emergency plane landing]).

Subjects who watched the video answered these questions immediately after it ended; subjects who did not watch the video answered these questions immediately after the instruction to imagine the scenario. Because we were interested in people's

quick, “gut feel or hunch,” subjects who did not see the video answered these questions immediately, rather than at a delay matched to the duration of the video.

Results and Discussion

Recall our primary research question was: to what extent does watching a video of an expert performing a skill inflate people’s confidence in their own ability to perform the skill? Before answering this question, we first excluded 27 subjects who provided nonsensical descriptions of the situation they were asked to imagine (n = 1), failed our attention check (n = 6), had a valid pilot's license (n = 4), or had flown or landed a plane before (n = 22). We therefore retained 198 subjects in the final dataset.

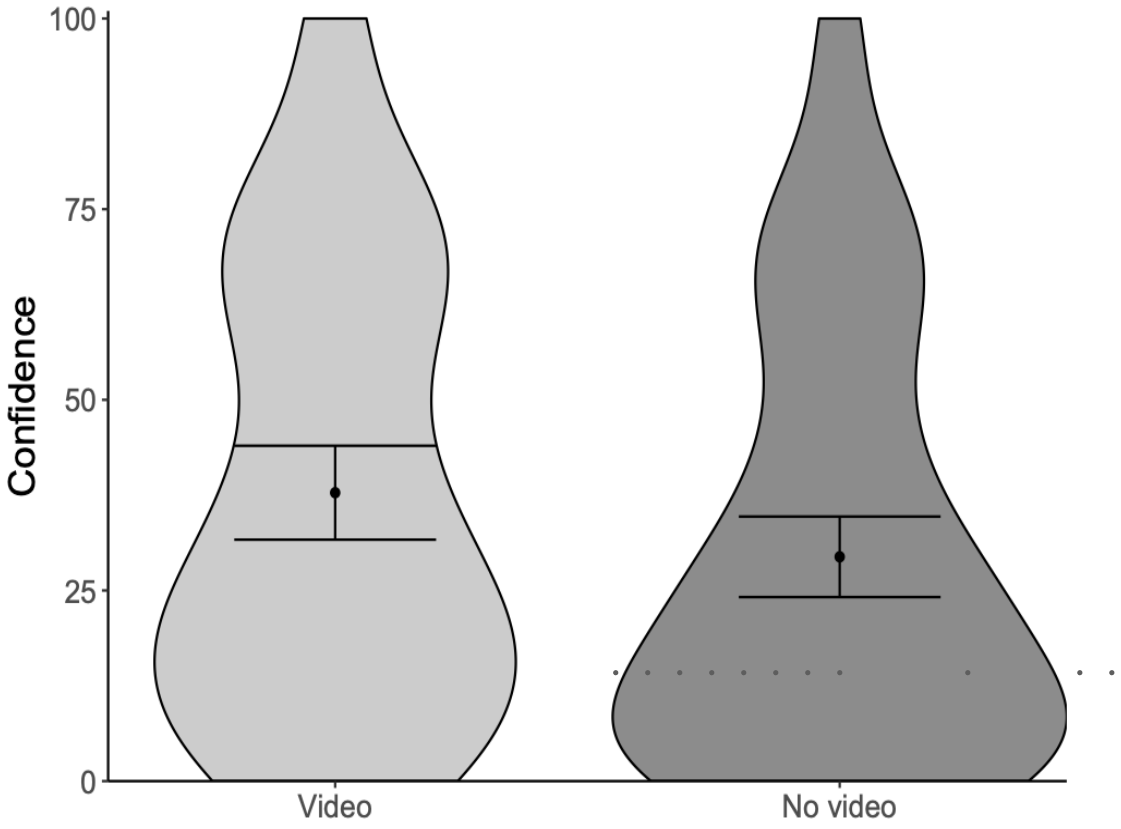
We now return to our primary question. To answer this question, we first classified subjects’ confidence ratings for the lower standard “without dying” question according to whether they had or had not seen the video.

We observed a marked skew in confidence ratings. The lower quartile was 9 and the modal response was 0; clearly, no transformation could restore normality. But t-tests—especially Welch’s, a parametric test that does not assume equal variance—are robust to most deviations in normality (R; Fagerland & Sandvik, 2009; see also Delacre et al., 2017 for the Welch’s test as a default). Therefore, we carried out a Welch’s t-test. Nonetheless, we also carried out a Wilcoxon signed rank test, a non-parametric test. Both of these analyses revealed the same pattern, and we display these findings in Figure 2. The figure shows two important findings. First, compared with subjects who had not watched the video, subjects who had watched it were more confident they could land the plane without dying, M_{video} = 37.82, M_{novideo} = 29.41, M_{diff} = 8.41, 95% CI [0.35, 16.46]. In null hypothesis testing terms, both the Welch’s t-test and Wilcoxon signed rank test showed the same pattern, $t(184) = 2.057$, $p = .041$; $Z = 2.15$, $p = .032$. Second, we asked a pilot with 35 years’ experience flying commercially and

internationally for Air New Zealand for his estimate on the chances of surviving the landing. His generous estimate was 10-15%—the dotted line on the figure represents this estimate.

Figure 2.

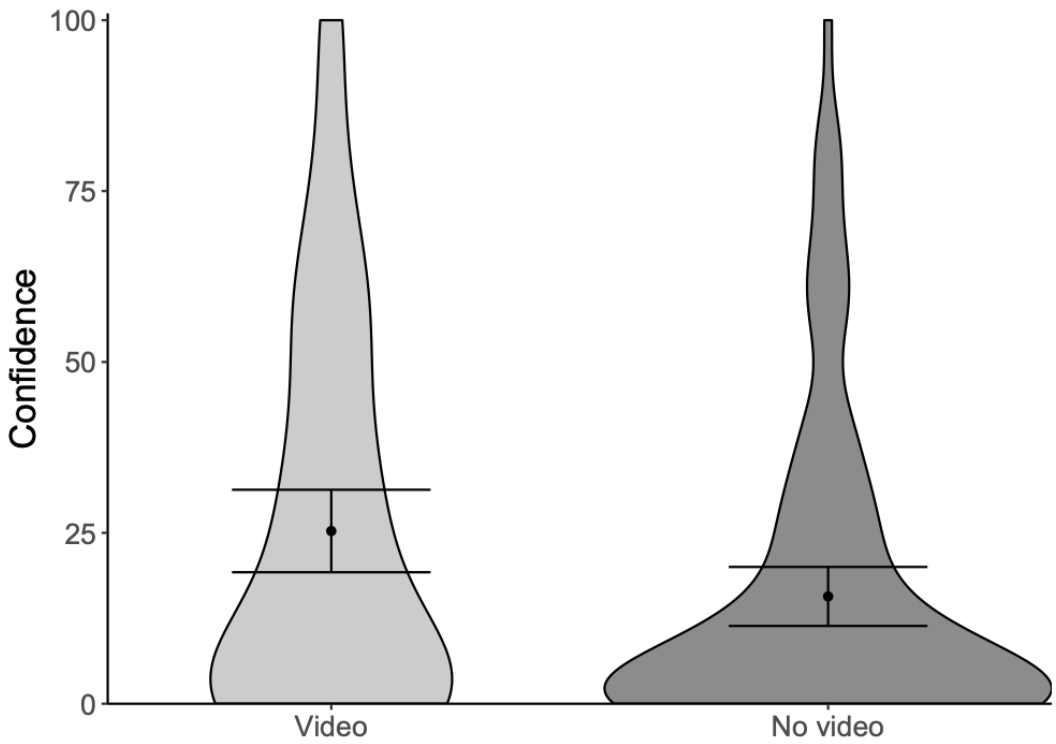
Subjects' confidence ratings for the lower standard "without dying" question by condition: video, no video.



Note. Error bars represent 95% confidence intervals of the cell means.

We then carried out the same analyses on subjects' confidence ratings for the higher standard—the "as well as a pilot could" question—and we display these data in Figure 3. As the figure shows, we found the same pattern as with the "lower standard" question. That is, subjects who watched the video were more confident than those who did not, $M_{video} = 25.27$, $M_{novideo} = 15.70$, $M_{diff} = 9.57$, 95% CI [2.20, 16.95]. The Welch's t-test and Wilcoxon signed rank test showed the same pattern, $t(166) = 2.56$, $p = .011$; $Z = 2.15$, $p = .031$.

Figure 3
Subjects' confidence ratings for the higher standard "as well as a pilot could" question by condition: video, no video.



Note. Error bars represent 95% confidence intervals of the cell means.

Our next step was to determine if men and women reported different degrees of confidence. To address this issue, we classified responses by gender and video condition, and calculated the mean confidence ratings. Collapsed across all conditions, men were more confident than women, M_{diff} = 12.24, 95% CI [2.39, 22.10], $F(2, 197)$ = 4.41, p = .01.

Next as a preliminary investigation of mechanism, we determined the extent to which people who watched the video of the pilot land the plane reported it was easier to imagine themselves landing a plane, relative to their counterparts who did not watch the video. We found no evidence to support this idea (M_{video} = 3.00, $M_{novideo}$ = 3.02; M_{diff} = -0.02, 95% CI [-.42, 0.39], p = .93). But in exploratory follow-up analyses, we found people's ease of imagining and confidence they could land a plane without dying were strongly associated when they had watched the video ($r(87)$ = -.57, 95% CI [-.70, -.41],

$p < .001$) and weakly associated—including plausibly zero—when they had not ($r(107) = -.18$, 95% CI [-.36, .01], $p = .06$). A Fisher's z test indicated that these correlations were significantly different ($r_{diff} = -.39$, 95%CI [-.61, -.15], $z = -3.18$, $p < .001$). Considered together, these findings fit with the idea that when people watched the video of the pilot land the plane, they were more sensitive to ease of processing cues (Schwarz, 2012). Moreover, the video might enable subjects to develop more detailed imaginations of themselves landing the plane, that they misconstrue as evidence that they could actually land the plane.

Taken together, these findings show that watching one short non instructional video of a pilot landing a plane inflated people's confidence in their own ability to land a plane—a specialized task requiring a great deal of expertise. Of course, one counterexplanation for our results is that people do not know landing a plane requires much expertise, or that after watching the video, people reevaluate their knowledge about what expertise is necessary. To address this possibility, we classified subjects according to whether or not they had watched the video, and then calculated their mean response to the question "How much expertise do you think is involved in landing a plane?" (1 = *None*, 5 = *A great deal*). These means approached ceiling and were nearly identical; people recognised that a great deal of expertise was necessary to land a plane, and this recognition did not differ as a function of whether people watched the video, $M_{video} = 4.39$, $M_{novideo} = 4.36$, $M_{diff} = 0.03$, 95% CI [-0.30, 0.37], $p = .83$. Considered together, these data do not support the counterexplanation.

The data from Experiment 1 provide preliminary evidence that increasing semantic context can inflate people's confidence in their own ability to perform a specialized task. In Experiment 2 we aimed to replicate these findings with a larger sample size to obtain a more precise estimate of the size of the effect (Cumming, 2012).

Experiment 2

Method

Subjects. We recruited subjects from Amazon's Mechanical Turk (MTurk), an online source of a vast and diverse population, who received $0.30 USD for participating (see Mason & Suri, 2012). We use the Shiny Web "power for two independent groups t-test" app (https://designingexperiments.com/shiny-r-web-apps/) to calculate sample size based on the data from Experiment 1. Using those data and a desired power 0.90, the target *n* per condition was 226. Anticipating exclusions, we aimed to collect data from 600 people, but because Qualtrics and MTurk interact in such way that it is possible to unintentionally collect more data than required, we ultimately collected data from a total of 632 subjects. After applying exclusions as detailed below, we retained 582 subjects in the dataset (M_{age} = 39.19, SD_{age} = 12.95; 31% identified as men, 68% identified as women, and 1% identified as gender diverse); 94% of subjects reported English was their first language.

Design. We used the same design as in Experiment 1, a between-subjects design with two conditions (Video: video, no video).

Procedure. We used the same basic method as in Experiment 1, but this time we counterbalanced the order of the "without dying" and "as well as a pilot could" questions.

Results and Discussion

Consistent with our preregistration, we first excluded 50 subjects who: provided nonsensical descriptions of the situation they were asked to imagine (*n* = 2), failed our attention check (*n* = 10), had a valid pilot's license (*n* = 8), or had flown or landed a plane before (*n* = 41). We pre-registered a paired comparison collapsing the order of the dying and pilot questions, because we did not expect to see an effect for order. When

we carried out this pre-registered comparison, we found subjects who watched the video of a pilot landing a plane were more confident in their own ability to land the plane at the lower standard, "without dying," $M_{video} = 34.23$, $M_{novideo} = 28.79$, $M_{diff} = 5.44$, 95% CI [0.85, 10.03], $t(561.43) = 2.33$, $p = .020$. But we did not find the same pattern for the higher standard, "as well as a pilot could," $M_{video} = 22.42$, $M_{novideo} = 19.20$, $M_{diff} = 3.22$, 95% CI [-.97, 7.42], $t(568.18) = 1.51$, $p = .13$.

When we took a closer look at the data using exploratory analyses, we found evidence that the order mattered as Figure 3 shows. That is, the left side of the figure shows that when subjects first saw the lower standard "without dying" and then the higher standard "as well as a pilot could" questions, the video boosted confidence on both. This condition was a replication of the method in Experiment 1. By contrast, the right side of the figure shows that when subjects first answered the higher standard and then the lower standard question, the video did not boost confidence on either measure.

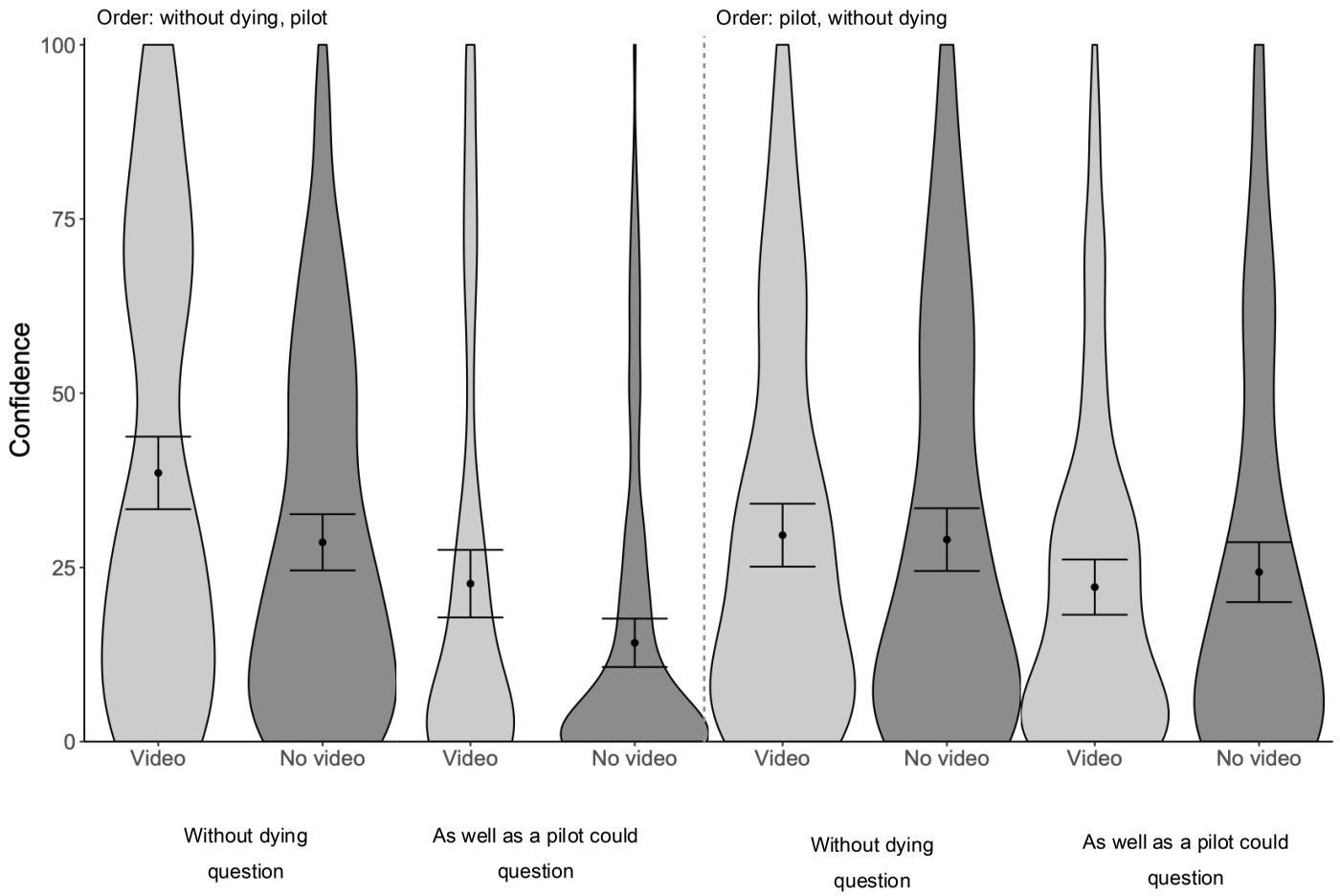
In null hypothesis testing terms, we conducted two separate 2 (Question order: dying first, pilot first) x 2 (Video condition: video, no video) ANOVA—the first on subjects "without dying" ratings and the second on their "as well as a pilot could" ratings. For the "without dying" ratings, we found a main effect of video condition, $F(1, 581) = 5.23$, $p = .02$; a marginal effect of question order, $F(1, 581) = 3.40$, $p = .07$; and an interaction, $F(1, 581) = 4.04$, $p = .04$, such that watching the video inflated people's confidence when the "without dying" question was asked first, $M_{diff} = 9.96$, 95% CI [1.60, 18.31], $p = .01$; but not when the "as well as pilot could" question was asked first, $M_{diff} = 0.64$, 95% CI [-7.90, 9.17], $p = .99$. For the "as well as a pilot could" ratings we found a main effect of question order, $F(1, 581) = 5.2$, $p = .02$, and an interaction, $F(1, 581) = 6.34$, $p = .01$, such that watching the video inflated people's confidence when the "without dying" question was asked first $M_{diff} = 8.49$, 95% CI [.87,

16.11], $p = .02$; but not when the “as well as pilot could” question was asked first, $M_{diff} = -2.15$, 95% CI [-9.93,5.63], $p = .89$.

The data showed small deviations from normality and variance. But when Blanca et al. carried out a series of Monte Carlo simulations with a range of distributions including extreme skewness and Kurtosis values (of up to skewness = 2 and kurtosis = 6), they found that ANOVA can be robust to such deviations (Blanca et al., 2017). Our skewness and kurtosis values were within these limits. Nonetheless, we also conducted a second analysis, a robust ANOVA (Hettmansperger & McKean, 2011). Using this approach, we found similar results as with the original ANOVA. More specifically, subjects who watched the video of a pilot landing a plane were more confident that they could land the plane without dying than people who didn’t; $F(1, 581) = 6.44$, $p = .01$. The order of the questions mattered such that people were more confident they could land the plane without dying when they were asked that question first; $F(1, 578) = 5.02$, $p = .03$. We also found a marginally significant interaction; $F(1, 581) = 3.53$, $p = .06$, such that watching the video inflated people's confidence when the "without dying" question was asked first.

Figure 4

Subjects' responses to the lower standard "without dying" and higher standard "as well as a pilot could" confidence questions split by condition (video, no video) and the order in which they were presented.



Note. Error bars represent 95% confidence intervals of the cell means.

Though these patterns are the result of exploratory analyses, they suggest the possibility that responses to the first question anchored responses to the second question—in particular, being asked the lower standard "without dying" question first allowed subjects to nod along with feelings of ease they experience through watching the video, and although the second question about the pilot standard might have shifted subjects away from their intuitive hunch, they adjusted insufficiently from their answer to the "without dying" standard (Kahneman, 2011; Tversky and Kahneman, 1974). We could recast this idea in "System 1 vs System 2" language. If so, we would say that when people considered their ability to perform the skill to a low standard, they took a "System 1," more heuristic approach, nodding along with recent feelings of easier mental simulation. But when other people compared themselves to a pilot, they took a "System 2," more systematic approach — because the comparison highlighted the expertise involved in performing the task. This distinction is broadly consistent with other approaches such as the SMF.

Next, and consistent with our pre-registration, we determined if men and women reported different degrees of confidence. To address this issue, we classified responses by gender and video condition, and calculated the mean confidence ratings. Collapsed across all conditions, subjects who watched the video were more confident than subjects who did not watch the video, $F(1, 578) = 6.59$, $p = .01$; and more to the point, men were more confident than women, $F(2, 578) = 9.21$, $p < .001$. Of course, these data are exploratory, not randomized, and we therefore urge caution in interpretation.

Then, and also as a pre-registered analysis, we determined if people who watched the pilot land the plane reported it was easier to imagine landing a plane relative to their counterparts who did not watch the pilot. As in Experiment 1, we found no evidence to support this idea when comparing mean responses to this question across the video and

no video conditions (M_{video} = 3.36, $M_{novideo}$ = 3.26; M_{diff} = 0.11, 95% CI [-.13, 0.34], p = .11). But, consistent with the pattern of results from Experiment 1, we found that people's ease of imagining and confidence they could land a plane without dying were moderately associated when they had watched the pilot ($r(279)$ = .43, 95% CI [-.52, -.33], p <.001) and weakly associated when they had not ($r(299)$ = 0.26, 95% CI [-.37, -.15], p <.001). A Fisher's z test indicated that these correlations were significantly different (r_{diff} = -.17, 95% CI [-.31, -.02], z = -2.27, p = .02). These findings provide more support for the idea that the video may enable subjects to develop more detailed imaginations of themselves landing the plane, that they misconstrue as evidence that they could actually land the plane.

Finally—and as in Experiment 1—we found that regardless of whether they saw the video, subjects reported similar awareness that being a pilot demands great expertise. That is, M_{video} = 4.50, $M_{novideo}$ = 4.49, M_{diff} = 0.01, 95% CI [-0.29, 0.31], p = .94.

Considered together, these findings show that when people watch a video of a pilot land a plane, they become more confident they could land a plane without dying. This finding replicates the major finding in Experiment 1. We also replicated the finding that men were more confident than women, and that people retained awareness of the fact that landing a plane requires expertise. Finally, our data add to the literature showing that when items tap into what people know, the order in which those items appear changes how they evaluate their prior performance (Weinstein and Roediger, 2012). Here, we extend this literature with data suggesting that when questions tap into what skills people think they have (but don't), the order in which those questions appear changes how they evaluate their future performance.

Still, we found no evidence to support the idea that watching the video made it easier for people to imagine landing a plane when comparing mean responses across the video and no video conditions. We expected that the video would give subjects more semantic context, making it easier to imagine themselves landing the plane. But one possible explanation is that we asked the "difficult to imagine" question after we had asked the two confidence questions. It is reasonable to speculate that the confidence questions themselves encouraged people to imagine landing a plane, and therefore contaminated people's responses to the imagination question. It is also possible that in asking people to consider the scenario, then watch the video or not, then answer questions about their confidence—and then assess how difficult it was to imagine themselves landing a plane—too much time had passed for people to answer the question accurately. To address these issues, we therefore conducted two more experiments in which, instead of asking subjects to report their confidence that they could land a plane, we instead asked them to rate how difficult it was to imagine landing the plane (in Experiment 3a) and how easy it was (in Experiment 3b).

Experiments 3a and 3b

Method

Subjects. A total of 206 people from Mechanical Turk took part in Experiment 3a, and 202 in Experiment 3b. In the absence of good evidence on which to project the size of the effect, we instead aimed for 100 people per cell.

Design. We used a between-subjects design with two conditions (Video: video, no video).

Procedure. In Experiment 3a, in line with our previous experiments, we asked subjects to "Imagine you are on a small commuter plane. Due to an emergency, the pilot is incapacitated and you are the only person left to operate the plane." Subjects were then randomly assigned to watch either a video of a pilot landing a plane, or to watch no such video. We then asked "How difficult was it for you to imagine yourself attempting to land the plane?" (1 = *Not at all*, 5 = *Very*).

In Experiment 3b, we asked subjects to perform the identical task, except we substituted the word "difficult" with the word "easy." "How easy was it for you to imagine yourself attempting to land the plane?" (1 = *Very easy*, 5 = *Not at all easy*).

Results and Discussion

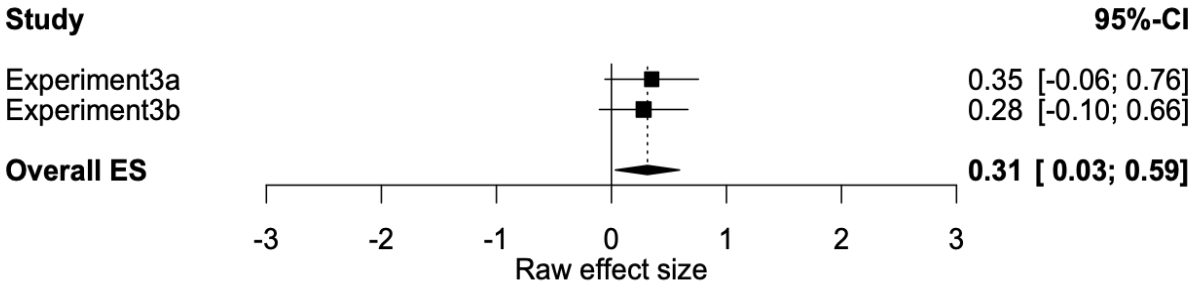
We applied the exclusions from the prior experiments. The final datasets therefore comprised 186 people in Experiment 3a and 176 in Experiment 3b.

We found similar results across both Experiment 3a and 3b. More specifically, in Experiment 3a, subjects who watched the video rated it directionally less difficult to imagine landing the plane than subjects who did not watch the video. In Experiment 3b, subjects who watched the video rated it directionally easier to imagine landing the plane than subjects who did not watch the video. That is, Experiment 3a ($M_{\text{video}} = 3.25$, $M_{\text{novideo}} = 3.60$, $M_{\text{diff}} = .35$, 95%CI [-.06, .76], $p = .09$; Experiment 3b ($M_{\text{video}} = 3.14$,

$M_{novideo}$ = 3.42, M_{diff} = 0.28, 95%CI [-.10, .67], p = .15. We then conducted a mini meta-analysis to obtain a more precise estimate of the effect (Cumming, 2012), and report those results in Figure 5.

Figure 5

Mini meta-analysis of subjects' imagination ratings by condition (R code; Carter & McCullough, 2014).



The vertical line in Figure 5 shows no effect for the video; the right side of the vertical line shows an effect in which the video made it easier for subjects to imagine attempting to land the plane. As the figure shows, subjects who watched the video found it easier to imagine landing the plane than subjects who didn't. That is, there was a raw effect size of M_{diff} = .31 [.03, .59], p = .03, or a 7.8% shift towards easier imagining.

If we consider Experiments 3a and 3b together, the findings fit with the idea that the added semantic context provided by the video might enable subjects to develop more detailed imaginations of themselves landing the plane, that they then misconstrue as evidence that they could actually land the plane. They are also consistent with the SMF, and related work, which suggests that future events are thought more likely when those events come to mind easily, are full of sensory detail and low on cognitive markers of effort (see Johnson et al., 1993; Lindsay, 2008; Schacter, Addis, & Buckner, 2007).

Table 1

Table showing confidence ratings by condition and question order for Experiments 1 and 2.

		Dying first						Pilot first					
		Video			No video			Video			No video		
Exp	*Rating*	*Mdn*	M	95% CI	*Mdn*	*M*	95% CI	*Mdn*	*M*	95% CI	*Mdn*	*M*	95% CI
1	Without dying	30	37.82	[31.66, 43.98]	21	29.50	[24.17, 34.83]	-	-	-	-	-	-
	As well as a pilot could	15	25.27	[19.24, 31.30]	5	15.75	[11.39, 20.11]	-	-	-	-	-	-
2	Without dying	30	38.56	[33.35, 43.77]	20	28.60	[24.56, 32.63]	23	29.63	[25.11, 34.14]	20	28.99	[24.48, 33.49]
	As well as a pilot could	10	22.67	[17.82, 27.51]	3	14.18	[10.70, 17.65]	16.5	22.17	[18.20, 26.13]	15	24.32	[20.02, 28.63]

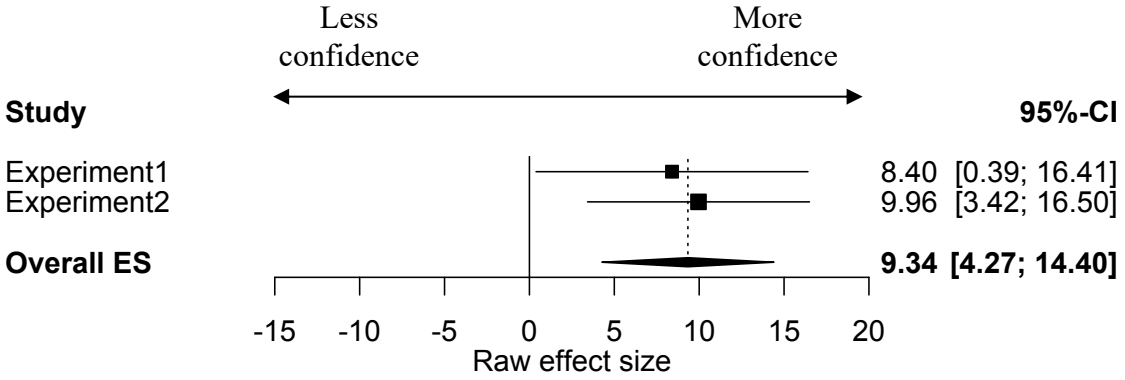
Chapter 3

Mini meta-analysis

To obtain a more precise estimate of the effect of watching a video on overconfidence in one's ability to perform a complex task, we conducted a mini meta-analysis of the data from Experiment 1 and the subset of the data from Experiment 2 that followed the same method (Cumming, 2012). We report the results in Figure 6.

Figure 6

Mini meta-analysis of subjects' confidence ratings by condition (R code; Carter & McCulloug 2014).



The vertical line in Figure 6 shows no effect for the video; the right side of the vertical line shows an effect in which the video made subjects more confident in their ability to land a plane without dying. As the Figure shows, subjects who watched the video were more confident in their own ability to land the plane than subjects who didn't. That is, there was a raw effect size of M_{diff} = 9.34 [4.27, 14.40], $p < .01$, or a 9.34% shift towards overconfidence.

These findings fit with the idea that watching a short video of expert performing a highly specialised task inflates people's confidence they too could perform the task.

How are we to understand this effect size? We can address this question by first turning to the related literature. First, our effect size is similar to that of the effect

called ‘truthiness,’ whereby photos boost the perceived truth of false claims—a mechanism thought to contribute to the ‘fake news’ problem (Newman et al., 2020). Second, our 9% shift in overconfidence is similar to that demonstrated in a recent study on ‘imagination inflation’, whereby imagining a hypothetical event increased people's certainty that the event really happened. This mechanism is thought to contribute to problems associated with recovered memory therapy (Calvillo et al., 2019). Third, our effect size is similar to a related illusion in which people who watched a video of a magician performing the tablecloth trick were more confident in their success actually performing the trick than people who merely thought about performing the trick but did not watch the video (Kardas & O’Brien, 2018). Fourth, our effect size is similar to another effect reported in the same paper in which people who watched a video of an expert throwing a dart at a bullseye, 20 times in a row, predicted they would score approximately 10 more points if they actually then threw a dart than people who watched the same expert video only once. This finding in particular highlights the pernicious effect of exposing people to multiple—but trivially informative—demonstrations of an expert performing a skill (Kardas & O’Brien, 2018). And finally, if we look to the wider literature, our effect size expressed as a correlation ($r = 0.124$) is reminiscent of the effect of antihistamine reducing runny noses, the link between combat exposure in Vietnam and PTSD over the next two decades, and the link between exposure to lead and impaired intelligence scores in children (Meyer et al., 2001.

Chapter 4

General Discussion

We found evidence that simply watching one non-instructional demonstration of an expert performing a highly complex skill led people to become more confident in their ability to perform that skill. More specifically, in Experiment 1, we found that watching a trivially-informative video of a pilot landing a plane inflated people's confidence in their own ability to land the plane—a specialized task—despite reporting that this skill requires a great deal of expertise. This conclusion fits with similar findings in the metacognitive literature on the various ways in which people use "feelings as information" (Schwarz, 2012). More specifically, people often interpret easy processing, or *fluency*, as (sometimes unwarranted) evidence of understanding (Schwarz, 2012; Alter & Oppenheimer, 2008; Cardwell, Lindsay & Garry, 2017).

In Experiment 2, we replicated the main findings that when people watch a video of a pilot land a plane, they are more confident they could land a plane without dying, whilst retaining their awareness of the fact that landing a plane requires expertise. But we found evidence that the order in which the confidence questions were asked mattered ("How confident are you that you would be able to land the plane without dying?" [0 = *Not at all confident*, 100= *Very confident*], and "How confident are you that you would be able to successfully land the plane as well as a pilot could?" [0 = *Not at all confident*, 100 = *Very confident*]). That is, when subjects first saw the lower standard "without dying" and then the higher standard "as well as a pilot could" questions, the video boosted confidence on both. But when subjects first answered the higher standard and then the lower standard question, the video did not boost confidence on either measure. These findings fit with an anchoring and adjustment explanation—being asked the lower standard "without dying" question first allows subjects to nod along with feelings of ease they experience through watching the video,

and although the second question about the pilot standard might shift subjects away from their intuitive hunch, they adjust insufficiently from their answer to the "without dying" standard.

In Experiments 3a and 3b, we found that subjects who watched the video found it easier to imagine landing the plane than subjects who didn't. These findings fit with the idea that the added semantic context provided by the video may enable subjects to develop more detailed imaginations of themselves landing the plane, that they misconstrue as evidence that they could actually land the plane.

Taken together, these findings support the idea that when people are in situations that encourage fluent processing, they may become overconfident in their assessments of what they know and can do.

Implications

These findings contribute to several literatures. First, our findings suggest that increased semantic context creates illusions of not just prior experience, knowing, or understanding—but the ability to actually do something implausible (Cardwell, Lindsay & Garry, 2017; Newman et al., 2012, 2015; Whittlesea, 1993). Our findings also extend prior work showing that increased semantic context leads people to predict rosy outcomes in the future. In one series of experiments, people evaluated positive or negative claims about the price of commodities in the future, such as manganese "is likely to have increased [decreased] in price three months from today" (Newman et al, 2018, p. 1225). Sometimes those claims appeared with a photo of the commodity; other times the claim appeared alone. Across those experiments, photos promoted rosiness about the future events, leading people to believe positive claims about the future but not negative claims. Here, we show that watching an expert perform a highly

specialized task promotes rosiness about people's ability to do something they almost certainly can't do.

Second, our findings fit with the literature suggesting that when imagining an experience in the future brings those self-generated mental products to mind easily, with low cognitive markers of effort, people think that experience is more likely to happen (Sherman et al., 1985; Newman et al., 2018, Gregory et al., 1982). Take, for example, one recent study in which people came to think there was a greater risk of experiencing climate change-related events when they conjured up easy-to-imagine hypothetical scenarios, such as the roads flooding while they were driving, than difficult-to-imagine ones, such as swimming in a lake in which soaring temperatures have promoted the growth of dangerous bacteria (Bø & Wolff, 2020). Our exploratory analyses in Experiments 1 and 2 provide preliminary evidence for a similar mechanism: when people watched the pilot land the plane, there was a strong association between how easy they found it to imagine landing the plane and how confident they were they could the plane without dying. When people did not watch the pilot land the plane, there was plausibly no association between these measures. These findings provide tentative support for the idea that the video may enable subjects to develop more detailed imaginations of themselves landing the plane, that they misconstrue as evidence that they could actually land the plane. These analyses were exploratory and such a conclusion is purely speculative. But our findings form 3a and 3b support a similar mechanism: that when people watched the pilot successfully landing a plane, they found it easier to imagine themselves performing the same feat, and were more confident that they could land a plane themselves in the future. Considered together, these findings suggest imagination does not only increase people's estimates of the likelihood of future experiences, but it can also inflate their confidence in their abilities.

Third, these findings extend what we know about the kinds of situations in which people experience "knowledge neglect,"—the failure to retrieve or apply their previously-demonstrated knowledge, resulting in a failure to spot errors (Fazio et al., 2013; see Marsh & Umanath, 2014, for a review). Here, we provide evidence that people also fail to retrieve or apply their demonstrated knowledge about the expertise involved in executing a complex skill when evaluating their own ability to execute that skill. Knowledge neglect is thought to arise from two routes. The first is that people "transport" themselves into fictional stories or hypothetical scenarios to better enjoy that material (Gerrig, 1993). The idea is that transportation involves some suspension of disbelief, which suggests a suspension of monitoring knowledge about the world. This suspension of monitoring encourages knowledge neglect and persuasion in line with the theme of the story or scenario (Green & Brock, 2000). There is also support for the idea that knowledge neglect occurs because people tend to refrain from engaging in the extra effortful processing that would avoid that neglect (Shah & Oppenheimer, 2008). For example, when people first encounter information, they default to representing it as true, and discounting or disproving it requires an additional step of extra effortful processing (Gilbert, 1991). A similar line of work suggests people often make bad decisions because they rely on easily-accessible information instead of spending more effort to access better information (Kahneman & Frederick, 2002). One consequence of this idea is that when new information partly overlaps with prior knowledge, the inherent repeated elements might produce a "good enough" signal of truth for people to avoid the additional effortful step (Reder & Kusbit, 1991). Such a possibility fits with the idea that knowledge neglect is exacerbated when decisions are susceptible to fluency-driven biases (Brashier et al., 2020). Here, when subjects first considered the emergency scenario and then watched the pilot landing a plane, knowledge neglect may

have arisen from greater transportation into the scenario and easily accessible feelings of processing ease that people misattribute to their abilities.

Fourth, they fit with an anchoring-and-adjustment account in which the order of items affects people's evaluations of their performance on those items (Tversky & Kahneman, 1974). That is, our findings from Experiment 2 support the possibility that the question people encountered first anchored their response to the second. In other words, when people first face the lower standard 'without dying' question, they might be lulled into a more heuristic type of processing, essentially 'nodding along' with the scenario, and tending to rely on recent feelings of easy processing that the added semantic context of the video provided (Schwarz et al., 2016). Then, to the extent people have difficulty sufficiently adjusting away from their overconfidence, when they face the higher standard question—comparing themselves with a pilot—they report more confidence than people who did not watch the video. But what happens to people who encounter the questions in the opposite order? When they are asked to compare their ability with that of a pilot, the obvious high standard should make it difficult to nod along; when they are then asked about their ability in the lesser standard, they too adjust away from their low confidence—but never get to the point where they nod along. These are speculative mechanisms, and although they fit with the literature, more research is needed.

Although we cannot determine the extent to which these mechanisms, in whole or in part, drive our effects, across four experiments our data suggest a general mechanism in which the semantic context of the video helps people mentally manufacture thoughts, images and feelings consistent with the proposition that they could land the plane. But there is more work to be done. For instance, it is not clear exactly what aspects of the video led people to be overconfident. For instance, our video depicted a successful,

smooth, landing with no complications. Perhaps if people saw a challenging landing, it might draw their attention to their genuine lack of skill. Furthermore, the video was not designed to be instructional (recall that an experienced commercial pilot who has trained many pilots pronounced it "useless"). Perhaps if the pilot stepped through the specific actions necessary to land the plane, the shift to a more concrete level of construal would encourage people to evaluate their abilities more accurately (Alter et al., 2010). We know from work on the illusion of explanatory depth, that people's confidence in their knowledge of complicated processes decreases when they are asked to provide a step-by-step explanation of how the process works (Rozenblit & Keil, 2002). For instance, when people considered their understanding of how a zipper works, they reported greater understanding than when they were asked the question at a more concrete level, such as how the parts of a zipper enable it to work (Alter et al., 2010). Likewise, the relative lack of concrete steps in the video may have inflated people's confidence. On the flip side, to the extent prior work on truthiness illuminates some of the mechanisms driving our effects, it might be fruitful to introduce another condition in which some people saw an unrelated video. In the truthiness literature, people tend to classify more difficult trivia statements as false when those statements were paired with a semantically unrelated photo—compared with a related photo or no photo at all (Newman et al., 2015). These findings fit with the idea that the lack of semantic relatedness in the unrelated photos produced disfluent processing.

In addition, the successful, smooth landing with no complications left subjects in the dark about the decisions driving the pilot's actions; subjects may have developed their own error-filled notions to fill in the gaps. Such a process could be explained by the "Beginners bubble," (Sanchez & Dunning, 2018). The idea is that when novices learn just a small amount about a complex task, their confidence in their performance

disproportionately shoots up compared to their objective performance. The proposed mechanism for this bubble of overconfidence is that people develop incomplete or insufficient theories about how to perform the task, which in turn leads to rogue feelings of competence. Although our video was not intended to be instructional in any way, because we chose a highly specialized task with which people had no prior learning, it seems reasonable to speculate that people might have developed incomplete or insufficient ideas about how to land a plane.

Even though our findings fit with the idea that the video provided semantic context that supported fluent generation of relevant thoughts, images and feelings, it is possible that some mechanisms other than fluency have driven these effects. For instance, it is possible that our effects arise from the combination of knowledge neglect and the beginner's bubble. It is also worth considering the role of individual differences given that most people reported realistic appraisals of their ability to land a plane. First, the literature on the need for self-enhancement and narcissism suggests it can lead to 'overclaiming', or claiming to know terms that do not exist. Perhaps people high on these traits would be more susceptible to the illusions of confidence we report here (Paulhus et al., 2003; Paulhus & Harms, 2004). Second, the literature on imagination inflation suggests that people who are high on hypnotic suggestibility and dissociativity are prone to developing false beliefs they have experienced an event after imagining it. It is possible people high on these traits would be prone to the illusory confidence-boosting effects of semantic context (Heaps & Nash, 2001). This possibility fits with our findings, in both experiments, that people who found it easier to imagine the situation were more confident in their ability to land a plane. Third, work on the illusory truth effect suggests that people are more susceptible to fluency biases when they are high on the need for cognition; the idea is that elaborative thought enhances

processing and boosts familiarity at a later test (Newman et al., 2020). Considered together, these possibilities suggest that the role of individual differences in moderating our effects is a topic worthy of future research.

The effects we report here might also be considered new and surprising examples of the confidence-boosting effects of videos demonstrated by Kardas and O'Brien (2018), who found that when people watched (say) a video, 20 times, of a man throwing a dart into the bullseye, they were more confident they themselves could successfully throw a dart closer to the bullseye than people who saw the video only once. We set out to test the boundaries of such an effect. We thought that landing a plane would pose as a reasonable boundary for a few reasons. First, it is a highly specialized skill requiring years of expertise. Second, the consequences of failing to perform such a task successfully could have disastrous consequences. Our findings suggest that this phenomenon holds for even highly specialized skills—and further our understanding about the kinds of situations that might encourage people to be overconfident.

Our findings add to a growing body of research showing how relying on confidence as an indicator of understanding or ability can easily lead people to make mistakes (Kornell & Bjork, 2007; Persky & Schlesselman, 2020; Serra & Dunlosky, 2010; Schwartz & Metcalfe, 1992; Townsend & Heit, 2011). We provide evidence of rapid changes in confidence—for even highly-specialised skills—when no learning or skill acquisition has taken place. Our findings might also serve as a reminder for people to engage in more systematic processing when evaluating their knowledge and abilities, particularly in situations when learning something new, as people are more prone to overestimating their abilities when they are unskilled (Dunning, 2011). Although we know that overconfidence can be beneficial in many cases, it can also prove dangerous. Consider Do It Yourself (DIY)

projects for example—hundreds of people in New Zealand are injured each year because they attempt to perform DIY tasks that they lack the "know how" to complete (ACC, 2010; Masters, 2001). Our findings suggest people should pay close attention to the resources they use when teaching themselves something new; here, just one, short, uninformative, video boosted people's confidence. The more we know about how we can manipulate people's overconfident judgements in the moment, the better we can advise people about what to do when they experience feelings of ease with regard to a highly skilled task.

These rapidly acquired illusions of skill acquisition might also have implications for the workplace. We demonstrate how simply observing others complete tasks boosts our confidence in our own ability to perform that same task. Moreover, this inflated confidence is not indicative of any real learning (Kardas and O'Brien, 2018). Overconfidence developed in this way could lead people to feel ready to tackle problems and take on tasks, when in fact they are ill-prepared, resulting in poorer quality outputs.

But our findings also suggest at least one way to eliminate the illusion. When people first answered the higher standard "as well as a pilot could" question and then the lower standard "without dying" question, the video did not boost confidence on either measure. Put another way, the data are consistent with the idea that when people compared themselves to a pilot, they took a "System 2", more systematic approach — because the comparison highlighted the expertise involved in performing the task (Kahneman, 2011). If we can encourage people to adopt a more strategic processing style when evaluating their abilities, that moves people away from relying on heuristics, people's self-assessments might be better calibrated. For example, people's estimates about how much they know become more accurate when they are encouraged to make inferences about that knowledge during study (Glenberg & Epstein, 1985).

These implications also extend to related domains such as education. Semantic context is a common tool used by educators to promote understanding and recall when people are learning new information (Ausubel, 1960; Bransford & Johnson, 1972). Here, we provide evidence consistent with the idea that semantic context can also create illusions of confidence by making it easier for people to imagine themselves performing a task. If the semantic context we provided instead made it easier for people to comprehend information in the moment, like we see with many educational materials, would we see a similar illusion of overconfidence in people's estimates of how much they are learning? In other words, would people mistake easier processing, that arises from enhanced comprehension of information in the moment, as evidence their abilities would transfer to new situations? For example, when people are learning a new skill, semantic context may ease comprehension of new material—say, when learning a foreign language, watching videos with subtitles aids understanding by translating that language into one the viewer speaks. But people might misattribute this momentary feeling of ease understanding the video as evidence of more general understanding and develop an illusion of skill acquisition where it is not warranted. We examined this hypothesis further in Part 2.

Part 2

Chapter 1

While much of the world spent the last couple of years at home watching Netflix, a curious phenomenon emerged among some of us and our colleagues. When we watched foreign television shows or films, we seemed to lose awareness of the presence of subtitles and even came to think we were learning the language. But within seconds of turning off the subtitles, we realised we were not learning much at all. How are we to make sense of this illusion. Here we suggest subtitles provide "semantic context"— meaningful information that eases people's comprehension of new information, and helps them connect it to their prior knowledge (Bransford & Johnson, 1972). But work from several literatures suggests a possible dark side of semantic context, such that people misattribute that ease as evidence they could better understand a foreign language more generally. Across 7 experiments, we report evidence consistent with the idea that subtitles provide semantic context that can rapidly drive illusions of better understanding a foreign language.

Subtitles as an aid to understanding

Of course, it is widely demonstrated that subtitles can be helpful. Subtitles can improve attention, vocabulary, and written and verbal comprehension (see, for a review, Gernsbacher, 2015). Moreover, evidence from eye-tracking studies suggests people attend to subtitles automatically, with no tradeoff between the processing of the subtitles and the images in the film (Perego et al., 2009). One way to think about subtitles is that they provide semantic context. For example, consider the type of subtitles in which the audio track and the subtitles are both in your language. These "closed captions" help, for instance, when English-language speakers who have a hearing impairment watch Schitt's Creek; the subtitles provide the semantic context viewers can no longer pick up. Other times, the audio track and subtitles are both in a language you are trying to learn. In this case, subtitles can help people segment

the words being spoken, which in turn boosts understanding of the phonology of the foreign language (Bisson et al., 2014). And still other times, subtitles translate the foreign audio track into your language. Most of us who were glued to Squid Game followed the story because the subtitles provided meaning we could not get from the Korean audio track. Considered together subtitles can provide semantic context in various different ways to enhance comprehension and understanding.

Semantic context

Beyond subtitles, in the wider literature, evidence suggests semantic context boosts comprehension and understanding of information in the moment—by improving meaning and logic, helping people to bring related information to mind, and connecting new ideas to prior knowledge. For instance, titles, advance organizers, illustrations, or photos commonly provide semantic context and confer these benefits (Ausubel, 1960; Bransford & Johnson, 1972; Marcus et al., 1996; Mayer & Gallini, 1990). Take, for example, the classic experiments in which people heard a difficult-to-comprehend passage, deliberately constructed to make the meaning unclear (Bransford & Johnson, 1972). In one of these experiments, subjects listened to the following passage describing a common procedure:

> The procedure is actually quite simple. First you arrange things into different groups depending on their makeup. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo any particular endeavor. That is, it is better to do too few things at once than too many. In the short run this may not seem important, but complications from doing too many can easily arise. A mistake can be expensive as well. The manipulation of the appropriate mechanisms should be self-explanatory, and we need not dwell on it here. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to

> foresee any end to the necessity for this task in the immediate future, but then one never can tell. (Bransford & Johnson, 1972, pg. 722).

People who know upfront that this obfuscated passage is titled "washing clothes" understand and later recall it better than people who don't know the title (Bransford & Johnson, 1972). Related work shows people are better able to learn and retain new information when they are introduced to a higher-level concept early on. This "advanced organiser" serves as a framework to help people organise information that comes later and relate it to their existing cognitive structure (Ausubel, 1960). More specifically, when people read an article on an unfamiliar topic—for example, about the properties of steel—they retain that information better if the first paragraph describes the steel material in a more general, abstract, way that outlines the key features of the following paragraphs. Likewise, when people see pictures that illustrate the workings of a pump, they recall more conceptual knowledge about it (Mayer & Gallini, 1990). Effects such as these are thought to occur because semantic context helps people to fit new information into existing representations or hierarchical networks (Ausubel, 1960).

We see similar patterns in work on narrative structure. For instance, when semantic context clarifies the goals of characters in a story, people better comprehend the story and recall more of it later. In one study, for example, when people read about a farmer and his donkey, learning early on that the farmer was trying to get the donkey into its shed helped reveal the structure of all the subgoals and associated actions in the story and improved recall (Bower, 1978). Those who read the same story without knowing the farmer's goal had worse understanding and recall. In Bower's followup work, readers were better able to recall information in a story (say, the items on a young man's grocery list) when it was relevant to a character's goals (knowing that the man wants to gain weight for the football team). And, in one set of classic experiments, when people read about a house, its surroundings, and

contents, they remembered different aspects of the story depending on their own goal as a reader (Pichert & Anderson, 1977). So, readers asked to imagine they were burglars casing the house remembered the coin collection, while readers asked to imagine they were home-buyers remembered the leaking roof. These examples suggest that semantic context—operationalised here as information about goals—partitions the incoming information accordingly, helps people to apprehend or construct a structure, and to integrate information together into a coherent narrative.

Finally, we see similar effects in the literature on event perception. When people view unfolding events, knowing about an actor's goal helps them to predict what will happen next (see, for a review, Zacks, 2020). For instance, people's eyes look ahead to the next action as if they can predict the actor's future (Eisenberg et al. 2018). More specifically, when people know an actor's goal is to reach a certain object, they predict or anticipate the reaching—looking at the target object before the actor's hand even moves. Taken together, semantic context can help people better comprehend new information and recall it later.

The dark side of semantic context

Collectively, these literatures all converge on the idea that subtitles provide nothing but upside. But there could also be a dark side. Although subtitles ease comprehension of information in the moment, there is evidence to suggest people would draw on these feelings of easier comprehension to make what should be unrelated decisions. After all, we know that people draw on feelings of ease in the moment to inform all kinds of judgements—including familiarity, liking, and quality (Oppenheimer, 2008). Work from at least three other literatures fits with the idea that people are more likely to falsely recall information when it is processed fluently. First, when subjects read sentences constructed to help them predict a terminal word (*the stormy seas tossed the…)*, they were more likely to falsely remember having seen the word "boat" in an earlier phase than when a sentence was not predictive (*he*

saved up his money and bought a …; Whittlesea, 1993). The proposed mechanism for this finding is that people misattribute the easier processing of the word presented in a predictive context to familiarity, rather than semantic relatedness. Second, when people are presented with a list of words, they are more likely to falsely recall seeing a "lure" word if those words share similar semantic associations (Roediger and McDermott, 1995). The idea is that when people are presented with the initial list, each word in that list (a semantic associate of the lure) repetitively primes the lure making it more available in memory. In fact, evidence suggests that the fluent processing of these lures is enhanced as the list gets longer. Then, because the lure feels easier to bring to mind, people misattribute this fluent processing as familiarity, and falsely report having seen the word before (Roediger and McDermott, 1995). Third, when people viewed a series of clips of a woman making a sandwich, they were more likely to falsely remember seeing unseen clips if the clips they saw more clearly fit with their "sandwich making" schema (Gerrie et al., 2006). The idea here is that subjects who saw the events that fit with their schema would have found it easier to process and comprehend the story. And we know that people misattribute ease as evidence of familiarity, leading to a higher false recall. Taken together, these examples suggest that semantic context can ease processing of information, a feeling people misattribute to unrelated judgements such as familiarity.

There is also evidence that semantic context creates feelings of learning; both learning and the illusion of learning evoke positive feeling thought to influence subsequent judgements (Lakshmanan & Krishnan, 2011). For example, in one study people evaluated wines more positively when the wine label contained context that taught them a new concept—such as a wine called "Yellow Rick" accompanied by a picture of a yellow haystack (Cardwell, Newman, et al., 2017). Of course, learning that a rick is a haystack has no bearing on the quality of the wine, but in this experiment, the more confident people were

that they understood new words such as “rick,” the higher they rated the quality of the wine. Considered together, this work raises the possibility that subtitles could provide semantic context that encourages feelings of familiarity, prior exposure, and even illusions of learning a foreign language.

Future judgements

To the extent these possibilities are true, we might also see people become more confident they could use this newly-acquired foreign language in the future. Several lines of research from the cognitive literature support this possibility. First, increased semantic context encourages people to predict an optimistic future. In a series of experiments, people judged claims that certain commodities would increase in price as truer when they were paired with a semantically-related photo (Newman et al., 2018). Second, providing semantic context about a dietary supplement (say, an image of a heart that conveys its function) can increase people's evaluations about the benefits they will receive from consuming that supplement (Delivett et al, 2020). Third, semantic context can even create overconfidence about one's abilities. People have come to report distorted confidence they could accurately throw a dart, execute a complex magic trick, and as we demonstrated in Part 1, even land a plane in an emergency, after watching short videos of experts performing those skills (Kardas and O’Brien, 2018). These examples provide evidence that semantic context boosts self-evaluations of future performance. Related work from the metacognitive literature also supports the idea that semantic context that encourages subjective feelings of ease of comprehension might even produce illusions of learning a foreign language.

Metacognition

Metacognition—or thinking about thinking—encompasses the monitoring and control of cognitive processes during encoding, retention, and retrieval. For example, when learning or studying information, students’ monitor their learning and allocate study effort and time

accordingly. Of course, their ability to remember the content well depends on their ability to accurately monitor how much they know in order to allocate these resources appropriately. We can refer to the processes involving the assessment of our knowledge as "metamemory", and can divide metamemory judgements into two distinct categories: information-based (theory-based) or experience-based (mnemonic-based) (Koriat & Bjork, 2006; Koriat & Levy-Sadot, 1999). When making information-based judgments, people infer knowledge analytically, on the basis of explicit rules and theories they retrieve from memory, such as "my memory is better if tested immediately after study, instead of after a long delay" (Koriat & Bjork, 2006). Experience-based judgments on the other hand are implicit inferences people make about their future memory performance on the basis of various internal cues, such as processing fluency, that produce subjective feelings people associate with knowledge or learning (Koriat & Bjork, 2006). People often accurately assess their knowledge and future memory performance (Nelson & Dunloskey, 1991). But we know that metamemory can also be prone to errors—people's predictions about what they know and what they can do don't always align with their actual knowledge and abilities (Koriat & Levy-Sadot, 1999). We see evidence of these metacognitive illusions of knowledge or ability in the literature on judgements of learning (JOL). More specifically, when learners are asked to give a JOL and predict how well they will recall information in the future, their JOLs are often out of sync with their actual future performance (Ball et al., 2014; Benjamin et al., 1998; Castel et al., 2007; Kornell et al., 2011; Rhodes & Castel, 2009; see for a review, Soderstrom & Bjork, 2015). One common driver for these miscalibrated JOLS is ease of processing (Begg et al., 1989; Kornell et al., 2011; Koriat et al., 2004). For example, when people are presented with lists of words, they give higher JOLs to words presented in better visual clarity or a larger font (features that make the words feel subjectively easier to process), even though their recall is no better than words presented in poorer quality, or a smaller font (Rhodes & Castel,

2008; Yue et al., 2013). These findings suggest that when information is accompanied by feelings of easier processing, people might mistake that subjective feeling as evidence learning has taken place, even when it has not.

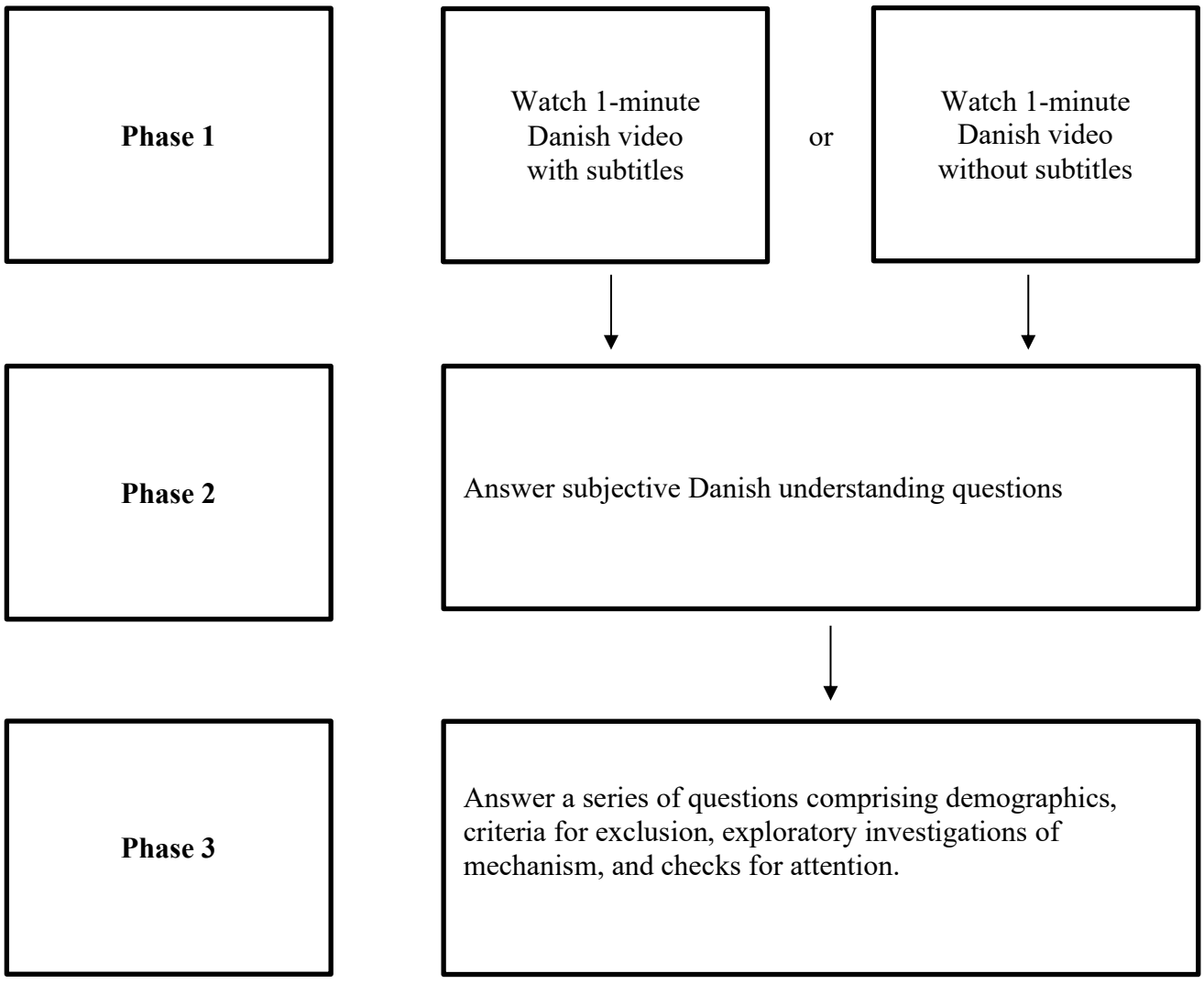
Overview of experiments

Taken together, these findings suggest that watching a subtitled, foreign-language video clip might enhance people's comprehension of information in the moment. But people might also mistake this ease of comprehension as evidence they could understand the foreign language more generally. So, when people who speak only English watch a short video clip from the Danish TV series "Rita," English subtitles should make it easier to comprehend the story—but people might also mistakenly interpret this ease of comprehension as evidence they would be better able to understand Danish in new situations. To examine these hypotheses, we asked this general question: To what extent would people who watched a subtitled Danish video clip become more confident in their ability to understand Danish, compared with those who watched the same video clip without subtitles? We addressed this question across a series of experiments by showing subjects a video clip of people speaking Danish. Some subjects saw the subtitled clip while others saw the unsubtitled version. Then we asked subjects to rate how well they thought they would be able to understand Danish in new situations. In Experiments 1a and 1b, we show that people who watched a brief, subtitled video from a Danish TV show thought they would understand a greater percentage of Danish dialogue if they were to see a new clip from the same episode of that show. In Experiment 2, we show a similar illusion, such that people who watched the subtitled video thought they would be better able to apply their newly-acquired comprehension of Danish to novel situations, compared to their counterparts who saw the video without subtitles. In Experiment 3a, we show this illusion is not dependent on a particular video, nor is it caused by people actually learning Danish, nor by people believing they could speak English instead of Danish

in the novel situations. In Experiment 3b, we found evidence supporting the idea that easier comprehension of information in the moment might be driving these illusions of foreign language ability; subjects who watched the clip with subtitles found it easier to comprehend and follow the story compared with subjects who watched the clip without subtitles. In Experiments 4a and 4b, we found the illusion persisted even after we asked people to listen only to the audio track from a video, thereby reducing social information. When taken as a whole, our results support the hypothesis that [a] semantic context boosts people's confidence by making it easier to understand the situation at hand, but [b] people then mistake that ease as evidence of their ability to comprehend a foreign language in novel situations.

Figure 1

Overview of the general method. The boxes summarise the key components across three phases.



Chapter 2

Experiment 1a

Our primary aim in Experiment 1a was to assess the efficacy of our materials and method.

Method

Subjects. We recruited subjects from Amazon's Mechanical Turk (MTurk), an online source of a diverse population (Mason & Suri, 2012). In the absence of data about the size of an effect, we aimed for 100 subjects for each condition, anticipating exclusions. But because MTurk and Qualtrics interact in such a way that it is possible to overcollect a target sample size, we collected data from 451 Mechanical Turk workers. We analysed the data from a total of 366 subjects after exclusions (M_{age} = 40.60, SD_{age} = 13.02; 33% identified as men, 66% identified as women, and 1% identified as gender diverse), and 94% of subjects indicated that English was their first language.

Design. We used a 2(subtitles: subtitles, no subtitles) x 2(clip origin: same episode, unspecified) between-subjects design. As preregistered, we had no interest in an interaction, our primary interest was in comparing subtitles and no subtitles within each scenario, rather than across scenarios. So, we compared the effect of subtitles vs no subtitles in two different ways. First, for people who were asked to predict their understanding of a subsequent scene in the same episode, we pre-registered a t-test for those who saw subtitles and those who did not. Second, for people who were asked to predict their understanding of another unspecified Danish video, we also pre-registered a t-test for those who saw subtitles and those who did not.

Procedure. First, subjects read about the task they would complete:

> With the rising popularity of online streaming services such as Netflix, people now have easy access to a wide range of media, including foreign films. Today you are going to watch a short clip from the Danish television series "Rita."

All subjects then watched a 1-minute video clip from “Rita.” In the scene, a teacher shows a new hire around a school, and everyone is speaking Danish. Half of the subjects were randomly assigned to see this clip with English subtitles; half saw it without.

Next, all subjects were told “Later on in this survey we are going to show you another short video clip, without subtitles, of people speaking Danish. This new video is 30-seconds long.” Then, half of all subjects were told “This new video clip is from the same episode of this TV show” while the other half did not receive additional information, thereby leaving the origin of the video unspecified. Then everyone was asked to predict“ How much of the spoken dialogue do you think you will understand in this video?" (0 = *None of the dialogue*, 100 = *All of the dialogue*). Next, everyone was asked “How much expertise do you think is involved in understanding a foreign language” (1 = *No expertise*, 5 = *A great deal of expertise*).

We then asked subjects a series of questions comprising checks for attention and compliance with the instructions, specifying that “As long as you complete the survey, we are going to pay you no matter what you tell us now, so please be honest. We need your honest answer so we know how to analyze the data you have provided us” (see supplemental for more information).

Results and Discussion

Recall our primary research question: To what extent would people who watched a subtitled Danish video clip become more confident in their ability to understand Danish, compared to those who watched the same video clip without subtitles? Before answering this question, we excluded from analysis the data from 85 (18.85%) participants who [a] failed our attention check/compliance measures; [b] provided a nonsensical description of the situation shown in the video; [c] had watched the television show “Rita” before; or [d]

reported speaking Danish. Across all exclusion criteria, 60.47% were excluded for failing one criterion, 3.49% for two, and 36.04% for more than two.

Next, we checked that people knew that understanding a foreign language requires a great deal of skill. In response to the question "How much expertise do you think is involved in understanding another language?" (1 = *No expertise*, 5 = *A great deal of expertise*), subjects reported that understanding a foreign language requires considerable expertise, and subtitle condition did not affect these ratings ($M_{\text{subtitles}}$=4.23, $M_{\text{nosubtitles}}$ = 4.09, M_{diff} = 0.14, 95%CI [-.07, .35], Welch's $t(319) = 1.29$ $p = .20$; Wilcoxon $Z = .93$, $p = .35$). That is, subjects knew that foreign language ability is a highly specialized skill.

We now return to our primary question. To answer this question, we first classified subjects' ratings of understanding according to whether they saw a video with subtitles or no subtitles, and then again by whether or not they were told the "new video clip is from the same episode of this TV show" before making their rating.

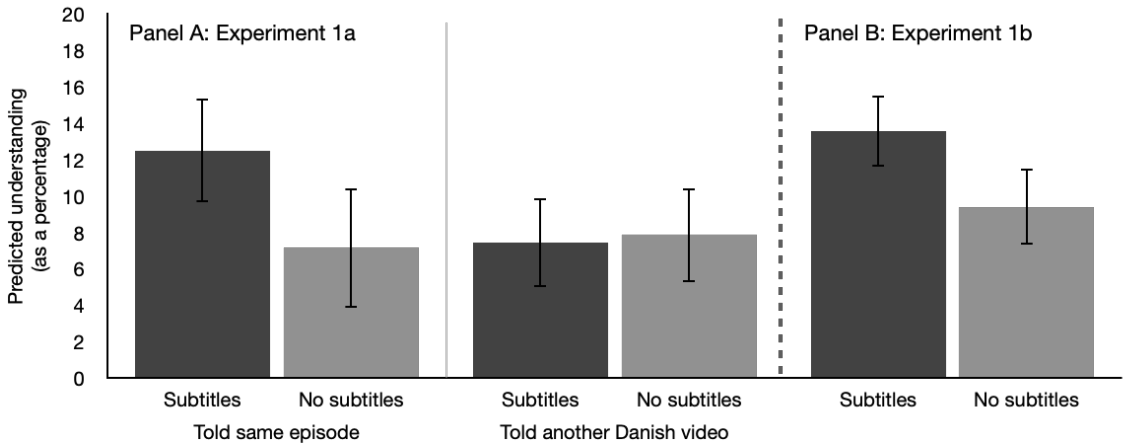
Recall that as we preregistered, we compared the effect of subtitles in two different ways. First, for people who were asked to predict their understanding of a subsequent scene in the same episode, we preregistered a *t*-test comparing those who saw subtitles with those who did not. Second, for people who were asked to predict their understanding of another unspecified Danish video, we also preregistered a similar *t*-test here.

We observed a marked skew in subjects' ratings of subjective understanding. The lower quartile was 1 and the modal response was 0; clearly, no transformation could restore normality. But t-tests are robust to most deviations in normality, especially Welch's, a parametric test that does not assume equal variance (Fagerland & Sandvik, 2009; Delacre et al., 2017). Therefore, we conducted a Welch's t-test. Nonetheless, we also conducted a non-parametric test, a Wilcoxon signed-rank test—to buttress our main analysis. Both analyses revealed the same pattern. We display the results in Figure 2, Panel A. As the left side of

Panel A shows, subjects who watched the video clip with subtitles predicted they would understand more Danish dialogue if they were to see a new clip from the same episode, compared with subjects who watched the clip without subtitles (M_{diff} = 5.34, 95%CI [0.99, 9.69]; Welch's $t(194) = 2.55$, $p = .02$; Wilcoxon $Z = 3.01$, $p = 0.01$). These findings fit with the idea that subtitles might boost people's comprehension of information in the moment and create illusions of skill.

Figure 2

S*ubjects' predicted Danish understanding by condition (subtitles, no subtitles) in Experiments 1a and 1b.*



Note. Panel A displays subjects' predicted Danish understanding ratings in Experiment 1a split by subtitle condition, and whether they were told the new clip would be from the same episode or not. Panel B displays subjects' predicted Danish understanding ratings in Experiment 1b where everyone was told the new clip would be from the same episode. Error bars represent 95% confidence intervals of cell means.

As the right side of Panel A shows, there was no evidence subjects who watched the clip with subtitles thought they would understand more Danish dialogue if they were to see a new clip of unspecified origin, compared with subjects who watched clip without subtitles (M_{diff} = -0.04, 95%CI [-3.96, 3.13], Welch's $t(164) = -.23$, $p = .41$;

Wilcoxon $Z = .73$, $p = .47$). These findings do not support the idea that comprehension of information necessarily creates illusions of skill in unspecified situations.

In an unplanned analysis, we ran a 2(subtitles: subtitles, no subtitles) x 2(clip origin: same episode, unspecified) between-subjects ANOVA on the data. We found an interaction, such that subtitles only inflated confidence when people were asked to consider watching another clip from the same episode, $F(1, 365)=3.92$, $p = .048$. But because this analysis is exploratory, we treat this conclusion as speculative.

The data from Experiment 1 provide preliminary evidence that watching a foreign-language video with subtitles inflates people's predictions of their ability to understand that language in similar situations, but not in unspecified situations. We might think of these two types of situations as providing “near” and “far” transfer, respectively (Royer, 1986; Schunk, 2012). These conclusions are speculative because we did not design Experiment 1a to compare across near and far transfer ratings. But because the data provide evidence only for near transfer, Experiment 1b focused on replicating only the “same episode” effect with a larger sample that would afford better estimation of the true size of the effect. We then return to the issue of far transfer in Experiment 2.

Experiment 1b

Method

Subjects. We collected data from 887 MTurk workers. After applying preregistered exclusions, 662 subjects remained in the dataset (M_{age} = 42.77, SD_{age} = 13.89; 33% identified as men, 65% identified as women, and 2% identified as gender diverse; 96% of subjects indicated that English was their first language).

Design. We used a two-group design manipulating subtitles (subtitles, no subtitles) between-subjects

Procedure. The procedure was identical to that of Experiment 1a, except with only the "near transfer" instructions. That is, when subjects were asked to predict how much of the spoken dialogue they would understand in the next video, they were all told the new video would be from the same episode of the TV show.

Results and Discussion

We first excluded, from the original dataset, data from 225 participants who [a] failed our attention checks (mostly people who failed to comply with instructions); [b] provided a nonsensical description of the situation shown in the video clip; [c] had watched the television show "Rita" before; or [d] reported speaking Danish. Across all exclusion criteria, 41% were excluded for failing one criterion; 12% for 2, and 47% for more than 2. We therefore retained a total sample of 662 subjects.

Recall our primary research question was: To what extent would people who watched a subtitled Danish video clip become more confident in their ability to understand Danish, compared with those who watched the same video clip without subtitles? To answer this question, we first classified subjects' understanding ratings according to whether they saw the video clip with or without subtitles, and display the results in Figure 2, Panel B. As the panel shows, we again found that subjects who watched the subtitled video of people speaking

Danish were more confident in their own ability to understand the dialogue in a new clip from the same episode, compared with people who watched the video without subtitles (M_{diff} = 4.17, 95%CI [1.38, 6.97], Welch's $t(639) = 2.93$, $p = .02$; Wilcoxon $Z = -1.96$, $p = .05$). These findings provide further support for the idea that subtitles inflate people's predictions of their ability to understand that language in similar situations.

But again—as in Experiment 1a—we found that subjects reported that understanding another language requires a great deal of expertise, regardless of whether they watched the video with subtitles or without subtitles ($M_{subtitles}$ = 4.08, $M_{nosubtitles}$ = 4.00, M_{diff} = 0.08, 95% CI [-0.09, 0.24], Welch's $t(649) = .97$, $p = .33$; Wilcoxon $Z = 1.26$, $p = .21$).

Experiment 2

In Experiment 2, we addressed the possibility that the reason we saw no evidence people thought they could carry out the "far transfer" task is not that it was a far transfer task, but that it was too abstract, and hard to imagine. For instance, asking people to consider seeing another video of people speaking Danish sets up an unconstrained future, in which the video could be almost anything—ice hockey, standup comedy, a Ministry of Finance press conference, or two people having coffee. We attempted to address this concern in two different ways, returning to "far transfer" tasks that were more constrained. More specifically, we asked people to imagine interacting with the people in the Danish video and then to rate their ability to [a] follow directions, and [b] make friends.

Method

Subjects. In the absence of relevant data about the size of an effect for the new "follow directions" and "make friends" dependent measures, we aimed for 150 subjects for each condition, anticipating exclusions. We collected data from 317 MTurk workers. After exclusions, we analysed the data from 295 subjects (M_{age} = 44.48, SD_{age} = 13.67; 33% identified as men, 67% identified as women). 95% of subjects indicated that English was their first language.

Design. We used a two-group design manipulating subtitles (subtitles, no subtitles) between-subjects.

Procedure. The procedure was same as that of Experiment 1a and 1b, with the exception of the key dependent measures. After subjects watched the clip, we asked them to "now insert yourself into this scenario and imagine interacting with these people." Then everyone answered two questions in a random order: " If one of these teachers were to give you directions to a classroom, how well would you be able to follow those directions?" and "If you were a new student at this school, how well would you be able to make friends?" To

increase fluent responding, we changed the scale so people considered the fluent response first (1 = *Very well*, 5 = *Not at all well*; Schwarz, 1999). But for ease of interpretation in this experiment, and in all others, we recoded and report the data to adhere to the convention in which higher numbers represent more of the construct being measured—here, better understanding of Danish.

Results and Discussion

We first excluded from analysis data from 22 people who [a] failed our compliance measures; [b] provided a nonsensical description of the situation shown in the video; [c] had watched the television show "Rita" before; or [d] reported speaking Danish. Across all exclusion criteria, 77% were excluded for failing one criterion, 18% for 2, and 5% for more than 2. We therefore analysed the data from 295 subjects.

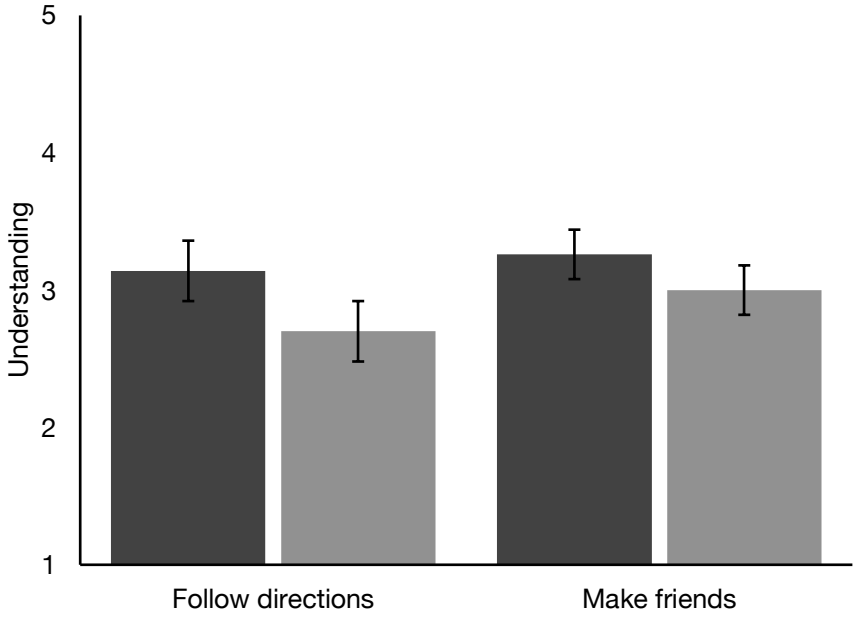
Recall our primary research question: To what extent does watching a subtitled video of people speaking a foreign language inflate people's judgments about how well they can understand that language in new situations? We followed the same approach as in Experiment 1a and 1b to classify and calculate the data. We display these results in Figure 3. As the figure shows, subjects who watched the video of people speaking Danish with subtitles thought they would be better able to follow directions and to make friends than people who watched the video without subtitles (Follow Directions: M_{diff} = 0.44, 95%CI [0.14, 0.74], Welch's $t(291) = 2.90$, $p = .01$; Wilcoxon $Z = -2.86$, $p = .01$; Make Friends: M_{diff} = .27, 95%CI [0.01, 0.53], Welch's $t(293) = 2.05$, $p = .04$; Wilcoxon $Z = 2.15$, $p = .03$). Put another way, if we consider the no subtitles group as people's baseline ability to understand Danish, watching the video with subtitles increased subjective understanding by 16.3%[4]. These findings provide further support for the idea that when subtitles ease

[4] We calculated this relative difference score via the following equation: ($M_{subtitles}$ - $M_{nosubtitles}$ / $M_{nosubtitles}$). Relative to the no subtitles group, the subtitles group rated their subjective Danish understanding 16.3% higher than the no-subtitles group.

comprehension of information, people use that ease as evidence they could understand a foreign language in new situations.

Figure 3

Subjects' mean ratings about how well they could "make friends" and "follow directions" in Danish.



Note. Error bars represent 95% confidence intervals of cell means.

As in Experiment 1a and 1b, regardless of whether they watched the video clip with or without subtitles subjects reported that understanding a foreign language requires much expertise, $M_{subtitles} = 4.18$, $M_{nosubtitles} = 4.01$, $M_{diff} = 0.17$, 95%CI [-0.05, 0.40], Welch's $t(293) = 1.53$, $p = .13$; Wilcoxon $Z = 1.89$, $p = .06$.

Considered together, these results suggest that watching a subtitled Danish video clip can inflate people's ratings of how well they will be able to understand Danish in future scenarios. Having said that, several obvious counterexplanations are worth addressing. First, it is possible people who see subtitles actually learn more Danish than those who don't, and so their inflated ratings of Danish understanding are justified. Therefore, to address this

counter-explanation, in Experiment 3 we gave everyone a short test of Danish vocabulary, at the end of the session.

Second, it could be that this effect is tied to the specific video. For instance, it might be that something about the friendly banter in the scene might encourage people to think they would thrive in Danish social settings. On a related note, it's possible that subjects who watched the video with subtitles were better able to make sense of the context of the school and the relationships between characters, and thus could draw inferences about following directions and making friends more than their no-subtitles counterparts. Therefore, to address this alternative explanation, in Experiment 3 we counterbalanced whether people saw the original Rita video, or another Danish video depicting a serious, chaotic political debate, and included dependent measures that were unrelated to the content in the videos.

Third, it is possible subjects thought they could communicate in English in the transfer situations. After all, the framing of the situation did not make it clear that everyone in the transfer situations would be speaking Danish. Of course, such a criticism does not explain the confidence-boosting effect of subtitles, but it might be introducing noise. Therefore, to address this counter-explanation, in Experiment 3 we made it clear that the people would be speaking Danish only in the transfer situations.

Experiment 3a

Method

Subjects. We used the Shiny Web app "power for two independent groups *t*-test" (https://designingexperiments.com/shiny-r-web-apps/) to calculate sample size based on the data from Experiment 2. Using those data and a desired power of .90, the target *n* per condition was 157, or a total of 314. We collected data from 357 workers before exclusions. After exclusions, we analysed the data from 308 subjects (M_{age} = 40.34, SD_{age} = 14.38; 36% identified as men, 63% identified as women, and 1% identified as gender diverse). 90% of subjects indicated that English was their first language.

Design. We used a between-subjects design with two conditions (subtitles: subtitles, no subtitles).

Procedure. First, subjects were randomly assigned to watch one of two videos of people speaking Danish, either with subtitles or without subtitles. One video was the "Rita" school tour used in Experiments 1a, 1b, and 2. The other video was from the Danish political drama "Borgen" and depicted several politicians engaged in a live, chaotic, televised debate. We counterbalanced the video that subjects saw. Subjects were then asked the following 4 questions, in a randomized order: "How well would you be able to follow Danish instructions in an emergency?", "If you were a new student at a Danish school and your peers spoke only Danish, how well would you be able to make friends?", "If you were listening to the Danish TV news, how well would you be able to understand the first story covered?", and "How well would you be able to understand the weather forecast on Danish TV?". All items were rated on 5-point likert scales (1 = *Quite well*, 5 = *Not at all well*). Next, everyone was asked "How much expertise do you think is involved in understanding a foreign language?" (1 = *No expertise*, 5 = *A great deal of expertise*). Subjects then listened to a series of Danish words

(17 in total; 5 that they had heard in the video, 5 from the video they didn't watch, and 7 others that are commonly used) and were asked to translate these words into English.[5]

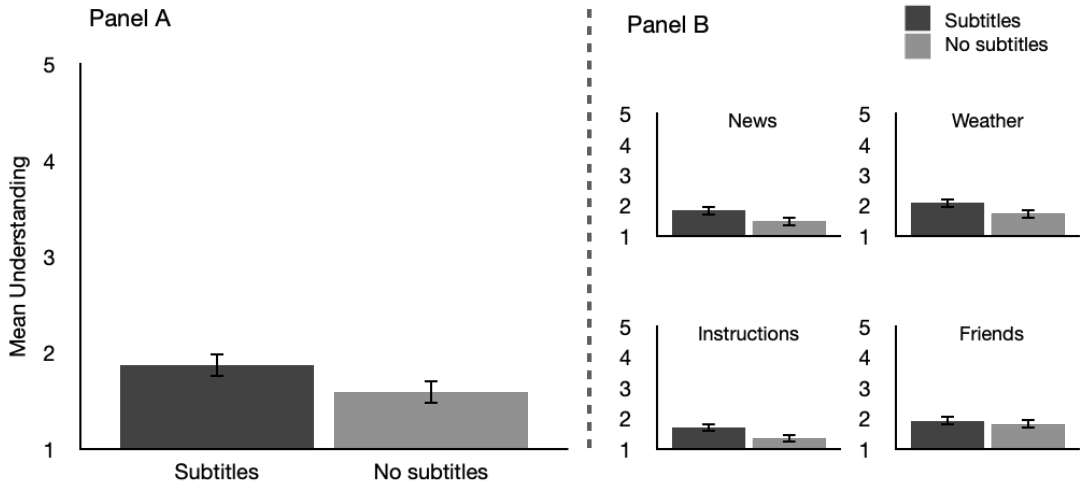
Results and Discussion

Planned Analyses. First, we excluded from analysis data from 49 participants who [a] failed our compliance measures; [b] provided a nonsensical description of the situation shown in the video; [e] had watched the television show "Rita" or "Borgen" before; or [f] reported speaking Danish. Across all exclusion criteria,72% were excluded for failing one criterion; 14% for 2, and 14% for more than 2. We calculated the mean of all four dependent measures for each subject. We then classified that mean according to whether or not subjects watched the video with subtitles. We display these findings in the main panel of Figure 4. As the figure shows, subjects who saw subtitles thought they would be better able to apply their newly-acquired comprehension of Danish to novel situations than those who didn't see subtitles, $M_{\text{subtitles}}$=1.87, $M_{\text{nosubtitles}}$=1.59, M_{diff} = .29, 95%CI [0.11, 0.46], Welch's $t(288)$ = 3.28, p = .01; Wilcoxon Z = 3.13, p = .01.

[5] We used the Lexiteria database to select commonly used words within the top 200 (https://lexiteria.com/word_frequency/danish_word_frequency_list.html). According to their website, "The Lexiteria Danish Word Frequency List 2010 contains 355,828 words taken from a 25,838,905 word corpus based on edited web pages in 2010."

Figure 4

Subjects' mean ratings about how well they would be able to understand Danish in new situations by condition (subtitles, no subtitles).



Note. Panel A displays subjects' mean ratings across the four dependent variables measuring subjects' subjective ability to understand Danish in new situations. Panel B displays the pattern for each of the individual items comprising that mean: News, Weather, Instruction, and Friends. Error bars represent 95% confidence intervals of cell means.

In a follow-up preregistered analysis, we repeated the analyses for each dependent measure considered individually. As Panel B in Figure 3 shows, subjects who saw subtitles thought they would be better able to follow Danish instructions in an emergency, understand the first story on the Danish TV news, and understand the weather forecast than subjects who didn't see subtitles (all reverse coded so higher bars indicate better understanding: Instructions: M_{diff} = 0.34, 95%CI [0.15, 0.53], Welch's $t(275) = 3.46$, $p = <.001$; Wilcoxon $Z = 3.39$, $p = <.001$; News: M_{diff} = .35, 95%CI [0.13, 0.57], Welch's $t(290) = 3.14$, $p = .01$; Wilcoxon $Z = 3.44$, $p = <.001$; Weather: M_{diff} = .34, 95%CI [0.11, 0.57], Welch's $t(297) = 2.85$, $p = .01$; Wilcoxon $Z = 2.84$, $p = .01$). There was no evidence subjects who saw

subtitles thought they would be better able to make friends than those who didn't (M_{diff} = 0.11, 95%CI [-0.13, 0.35], Welch's $t(303) = .93$, $p = .35$; Wilcoxon $Z = .86$, $p = .39$).

We then calculated subjects' mean percentage score on the translation test. Subjects performed poorly, regardless of subtitle condition ($M_{subtitles}$= 0.31%, $M_{nosubtitles}$ =0.30%, M_{diff} = 0.01%, 95%CI [-0.28, 0.30], Welch's t(306) = .08, $p = .94$; Wilcoxon $Z = .07$, $p = .94$). In other words, we found no evidence that people who watched the subtitled video were learning any Danish from the short video.

Finally, both those who saw subtitles and those who did not still rated the skill of understanding a foreign language as requiring considerable expertise, ($M_{subtitles}$ = 4.09, $M_{nosubtitles}$ = 4.19, Mdiff = -0.10, 95%CI [-0.32, 0.11], Welch's $t(303) = .94$, $p = .35$; Wilcoxon $Z = .88$, $p = .38$), a finding consistent with our prior experiments.

Considered together, the results from Experiment 3 provide further support for the hypothesis that subtitles enhance people's comprehension of information in the moment. But people mistake this ease of comprehension as evidence they could better understand the foreign language more generally, even though they didn't learn anything.

Unplanned Analyses. As part of routine analysis to assess new materials, we received feedback from colleagues that they found the new Borgen debate video harder to follow than the Rita school tour video. On the basis of this feedback, we examined our data to determine if the two different videos produced different mean understanding ratings. We carried out an unplanned 2(subtitles: subtitles, no subtitles) x 2(clip type: Rita, Borgen) analysis on subjects' mean understanding ratings. We found that it mattered which video subjects were randomly assigned to watch. That is, people who watched Rita thought they would understand more Danish than people who watched Borgen, $F(1, 304) = 4.61$, $p = .03$. But those saw the subtitled videos thought they would understand more Danish than those who

did not, $F(1, 304) = 10.59$, $p < .001$. There was no evidence of an interaction. Because this analysis was unplanned, we treat these results as tentative.

We then addressed these same issues a priori by collecting new data in Experiment 3a. More specifically, we wanted to know whether the lower understanding ratings for the new Borgen video were produced because that video was hard to follow.

Experiment 3b

Method

Subjects. We collected data from 138 MTurk workers. We analysed the data from 137 subjects after exclusions (M_{age} = 40.34, SD_{age} = 14.38; 15% identified as men, 82% identified as women, 3% identified as gender diverse; 96% indicated English was their primary language).

Design. We used a 2(subtitles: subtitles, no subtitles) x 2(clip type: “Rita”, “Borgen”) mixed design. Subtitles was a between-subjects factor and clip-type was a within-subjects factor.

Procedure. First, subjects were randomly assigned to watch one of the two videos from Experiment 3a, either with subtitles or without subtitles. One video was the "“Rita”" school tour; the other video was the "“Borgen”" chaotic political debate. Then subjects answered seven questions about how easy it was to understand and follow the storyline: “How clear did you find the story in the film?” (1 = *Not at all*, 7 = *Completely*), “How cohesive did you find the story in the film?” (1 = *Not at all*, 7 = *Completely*), “How coherent did you find the story in the film?” (1 = *Not at all*, 7 = *Completely*), “How well do you believe you understood what happened in the story?” (1 = *Not at all*, 7 = *Completely*), “Did the story in the film make sense to you?” (1 = *Not at all*, 7 = *Completely*), “As you watched the film, how easy was it for you to predict what would happen next?” (1 = *Not at all easy*, 7 = *Very easy*), “Did the sequence of events in the film have a logical flow?” (1 = *Not at all*, 7 = *Completely*). Then, subjects watched the other video (“Rita” or “Borgen”) and made all the same ratings.

Results

First, we excluded from analysis data from one participant who reported speaking Danish. We then addressed our primary question: to what extent did the type of clip people

watched, and whether the clip had subtitles, influence how easy it was to understand and follow the storyline? We report the results in two ways. First, because our primary interest was in how easy people would understand and follow the storyline in these two clips, we compared, between subjects, the ratings for the clip subjects saw first. We carried out 2 (subtitles, no subtitles) x 2("Rita", "Borgen") analyses on subjects' ratings for each of the 7 items. As shown in the Table 1, compared to subjects who watched the "Borgen" clip, subjects who watched the "Rita" clip found the video clearer ($F(1, 136) = 25.98$, $p < .001$), more cohesive ($F(1, 136) = 14.02$, $p < .001$), more coherent ($F(1, 136) = 39.20$, $p < .001$), they better understood what happened ($F(1, 136) = 32.34$, $p < .001$), the film made more sense ($F(1, 136) = 30.80$, $p < .001$) and had more of a logical flow ($F(1, 136) = 27.63$, $p < .001$) and it was easier to predict what would happen next ($F(1, 136) = 16.82$, $p < .001$). When we took the mean across these seven items, the data converge on the idea that the clip mattered; people who watched the Rita clip found it easier to follow the story than those who watched the Borgen clip, $F(1,136)=39.45$, $p < .001$.

Compared to subjects who watched the clip without subtitles, subjects who watched the clip with subtitles found the video clearer ($F(1, 136) = 20.81$, $p < .001$), more cohesive ($F(1, 136) = 13.02$, $p < .001$), more coherent ($F(1, 136) = 32.23$, $p < .001$), they better understood what happened ($F(1, 136) = 36.03$, $p < .001$), the film made more sense ($F(1, 136) = 38.03$, $p < .001$) and it was easier to predict what would happen next ($F(1, 136) = 9.83$, $p = .01$). They did not report that the clip had more of a logical flow ($F(1, 136) = .01$, $p = .92$). When we took the mean across these seven items, the data converge on the idea that subjects who watched the clip with subtitles found it easier to follow the story compared with subjects who watched the clip without subtitles ($F(1,136)=26.89$, $p < .001$).

Table 1

Subjects' mean ratings of how easy it was to understand and follow the storyline of the Rita and Borgen clips

	"Rita" clip						"Borgen" clip					
	Subtitles			No Subtitles			Subtitles			No Subtitles		
Rating	*Mdn*	M	95% CI	*Mdn*	*M*	95% CI	*Mdn*	*M*	95% CI	*Mdn*	*M*	95% CI
Clear	5.00	5.09	[4.36, 5.63]	4	3.62	[3.08,4.16]	4	3.48	[2.92, 4.05]	2	2.63	[2.20, 3.05]
Cohesive	5	4.97	[4.48, 5.46]	4	3.73	[3.28, 4.17]	4	3.70	[3.28, 4.12]	3	3.20	[2.62, 3.78]
Coherent	5	5.03	[4.52, 5.55]	4	3.68	[3.26, 4.09]	4	3.55	[3.08, 4.02]	2	2.37	[1.95, 2.79]
Understood what happened	5	5.25	[4.81, 5.69]	4	3.65	[3.09, 4.21]	4	3.73	[3.14, 4.32]	2	2.34	[1.95, 2.73]
Make sense	6	5.41	[4.87, 5.95]	4	3.59	[3.04, 4.14]	4	3.76	[3.15, 4.37]	2	2.31	[1.88, 2.75]
Predict	3.5	3.97	[3.47, 4.47]	3	3.27	[2.73, 3.81]	3	3.03	[3.47, 3.59]	2	2.17	[1.77, 2.58]
Logical flow	5	5.19	[4.75, 5.62]	5	5.14	[4.70, 5.57]	4	3.91	[3.41, 4.41]	4	3.91	[3.37, 4.46]

Second, to determine the relative ease with which people understood and followed the storyline, we compared, within-subjects, the mean understanding rating for both clips. Compared to subjects who watched the clip without subtitles, subjects who watched the clip with subtitles found it easier to understand and follow the storyline ($F(1, 136) = 36.36$, $p < .001$). Compared to subjects who watched the "Borgen" clip, subjects who watched the "Rita" clip found it easier to understand and follow the storyline ($F(1, 136) = 155.71$, $p < .001$). There was an interaction between mean understanding for each clip and which clip subjects saw first ($F(1, 136) = 22.79$, $p < .001$). Tukey's HSD tests indicated that subjects found it easier to follow and understand the "Rita" clip if they watched it after the "Borgen" clip relative to subjects who watched the "Rita" clip first, $M_{diff} = .98$ [.47, 1.50], $p < .001$; whereas subjects reported it was similarly easy to follow and understand the "Borgen" clip regardless of whether they saw that clip first or second $M_{diff} = .17$ [-.35, .70], $p = .82$.

If we step back now, these data support the idea that subtitles enhance people's comprehension of information in the moment, and people mistake this ease of comprehension as evidence they could understand the foreign language more generally.

Experiment 4a

Although our general research question concerns the extent to which semantic context boosts illusions of foreign language ability, in Experiments 1-3 we provide visual context—in addition to our subtitles manipulation—that we can't control. We know, for example, that nonverbal information such as body movement, facial expressions, and gesturing plays an important role in communication (Kellerman, 1992). The characters in our video clips gestured frequently, and we know that sociogestural information can convey meaning that observers reliably interpret (Goldin-Meadow & Alibali, 2013). In fact, second language learners' comprehension improves when they watch videos showing facial expressions and gesturing, relative to hearing foreign language audio alone (Sueyoshi & Hardison, 2005).

Figure 5

Example passages in Experiment 4.

Panel A	**Panel C**
Topic "Washing clothes"	**Topic "Washing clothes"**
The procedure is actually quite simple. First you arrange things into different groups depending on their makeup. Of course, one pile may be sufficient depending on how much there is. If you have to go somewhere else due to lack of facilities that is the next step. It is important not to overdo it. It is better to do too few things at once than too many.	Proceduren er egentlig ret simpel. Først organiserer du tingene i forskellige grupper afhængigt af deres udseende. Én bunke kan naturligvis være tilstrækkeligt, afhængigt af hvor meget der er. Hvis du bliver nødt til at tage et andet sted hen på grund af manglede faciliteter, er det det næste skridt. Det er vigtigt ikke at overgøre det.

This may seem unimportant, but complications from doing too many can easily arise. A mistake can be expensive. The manipulation of the appropriate mechanisms should be obvious. At first the whole procedure may seem complicated, but soon it will become just another facet of life. It is hard to foresee any end to the need for this task.	Det er bedre at ordne få ting på én gang fremfor for mange. Det kan virke trivielt, men der kan nemt opstå problemer, hvis man ordner for mange. En fejl kan være dyr. Mekanismerne og hvordan de påvirkes burde være indlysende. Til at begynde med kan hele proceduren virke kompliceret, men det bliver hurtigt bare endnu en del afl livet. Det er svært at forestille sig, at behovet for at udføre denne opgave forsvinder.
Panel B The procedure is actually quite simple. […] It is hard to foresee any end to the need for this task.	**Panel D** Proceduren er egentlig ret simpel. […] Det er svært at forestille sig, at behovet for at udføre denne opgave forsvinder.

In Experiment 4a we aimed to develop new materials that would minimize social information and eliminate sociogestural information. Therefore, we removed social information from the video clips; subjects could hear only the foreign audio while looking at a blank screen or a static picture. We adapted work by Bransford and Johnson (1972) and Wiley and Rayner (2000). More specifically, we adapted a difficult-to-comprehend passage, see for example Figure 5. Because this work is so well known, it is important to make clear

the significant ways in which our use of these materials was different. For instance, when Bransford and Johnson attached the title “washing clothes” to the passage in the left column of Figure 5, it provided semantic context and helped people scaffold incoming information relative to subjects who saw just the passage without a title. By contrast, we asked our subjects to listen to this passage translated to Danish, as the right column of Figure 5 shows. What then should be the effect of asking some people to listen to this Danish passage and telling them it was about "washing clothes”? Whereas, for Bransford and Johnson subjects, knowing the passage was about washing clothes made it easier to comprehend, for our subjects it could not. The only thing that knowledge would do for them is let them know the foreign words they are hearing are about washing clothes, as illustrated in Panel C.

And now consider what should be the effect of showing some people English language subtitles and not telling them what the passage was about. This situation essentially turns the passage into the “difficult-to-comprehend” control condition in Bransford and Johnson, in which people read a passage without being told what it was about—here equivalent to Panel B. On the basis of this thinking, we reasoned that only the combination of information about the topic plus subtitles would be equivalent to the experimental “easier-to-comprehend” condition in Bransford and Johnson, and produce easier comprehension. Remaining subjects who didn’t receive a title or subtitles, and simply listen to the passage spoken in Danish should experience even less comprehension than the “difficult-to-comprehend” condition in Bransford and Johnson because the passage is in Danish. Subjects’ experience, then, would be akin to turning on the radio to listen to the Danish news. Therefore, in Experiment 4a we addressed this question. To what extent would people who heard the Danish passage—knowing the topic or seeing English subtitles—mistake their ease of comprehension as evidence they could understand that language in new situations.

Before answering this question in Experiment 4a, we first normed the materials. We found these data support the hypothesis that, of the four conditions in Figure 5, only the "topic plus subtitles" condition was easy to comprehend.

Method

Subjects. We collected data from 333 MTurk workers, (on the basis of no prior work with which to estimate the size of the effect, we aimed for at least 75 subjects for each condition, anticipating exclusions.

Design. We used a 2(subtitles:subtitles, no subtitles) X 2(title:title and picture, no title or picture) between-subjects design.

Procedure. The experiment proceeded in three phases. In the first phase, subjects listened to one of three audio clips of people speaking a difficult-to-comprehend passage in Danish. Half the subjects were randomly assigned to listen to this passage while seeing English subtitles; at the same time these subjects were evenly divided between also seeing a picture (for instance, of a washing machine) and title ("washing clothes") or no picture or title. The other half of subjects were randomly assigned to listen to this passage without seeing English subtitles; again, half were evenly divided between also seeing a picture (for instance, of a washing machine) and title ("washing clothes") or no picture or title.

In the second phase, subjects answered the following four questions in a randomised order: "How well would you be able to follow Danish instructions in an emergency?" (1 = *Quite well*, 5 = *Not at all well*), "If you were a new student at a Danish school and your peers spoke only Danish, how well would you be able to make friends?" (1 = *Quite well*, 5 = *Not at all well*), "How well would you be able to understand the top story on the Danish TV news?" (1 = *Quite well*, 5 = *Not at all well*), and "How well would you be able to understand the weather forecast on Danish TV?" (1 = *Quite well*, 5 = *Not at all well*). Again, subjects were asked "How much expertise do you think is involved in understanding a foreign

language" (1 = No expertise, 5= A great deal of expertise) and asked to rate "How difficult was it to comprehend the video?" (1-*Very easy*, 7-*Very difficult*).

In the third phase, subjects listened to a series of Danish words (16 in total: 4 they had heard in the video, 6 from the videos they didn't watch, and 6 others that are commonly used) and were asked to translate these words into English.

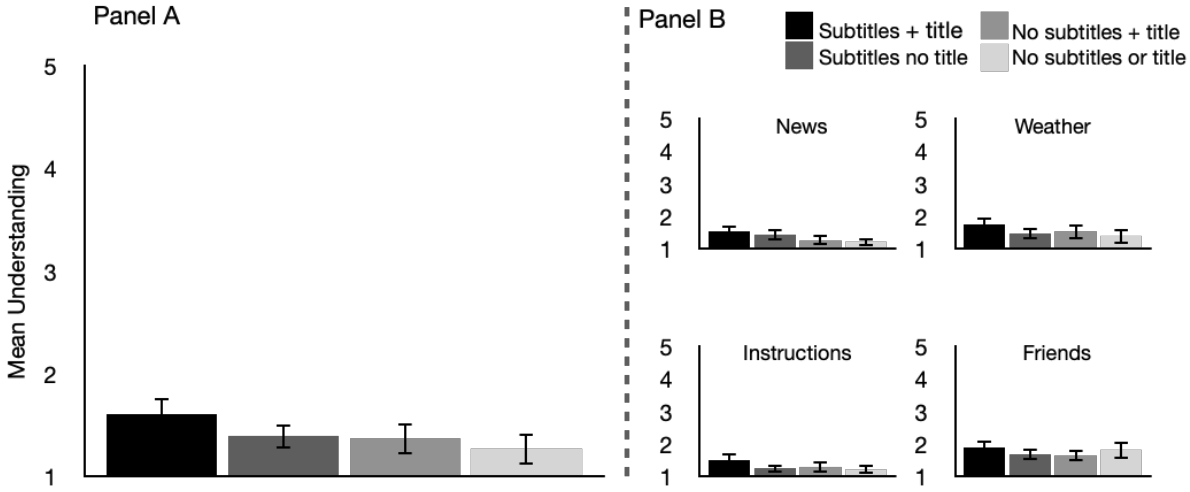
Results and Discussion

First, we excluded from analysis data from 19 participants who [a] failed our compliance measures; [b] provided nonsensical responses, or [c] reported speaking Danish. Across all exclusion criteria, 100% were excluded for failing one criterion. Next, we looked at the distributions of subjects' responses to the four main dependent variables. We noticed that the "friends" item had different patterns and upon further analysis, was less correlated with the other three confidence ratings. We also received comments from subjects suggesting this "friends" item was less reliant on foreign language understanding, such as: "There have been several occasions in my life when I have made friends with people who did not speak English, and whose language I did not speak." Therefore, we decided not to include the item in the mean confidence rating, and instead we carried out 2(subtitles, no subtitles) x2(title, no title) analyses on subjects' mean confidence ratings for the remaining three dependent variables: following Danish instructions, understanding the first story on the Danish TV news, and understanding the Danish weather forecast. We found that the context of the video mattered. That is, people who watched the video with a title were more confident than people who watched the untitled video, ($F(1, 313) = 5.18$, $p = .02$). But those saw the subtitled videos were more confident than those who did not ($F(1, 313) = 6.57$, $p = .01$). There was no evidence of an interaction. Pairwise comparisons indicated that compared to subjects who watched the Danish video with no subtitles or title, subjects who watched the video with subtitles and a title reported they were better able to understand Danish more generally

(Tukey's HSD effect size: M_{diff} = .28 [0.01, 0.52], p = .04). All other conditions did not significantly differ from one another.

Figure 6

Subjects' mean ratings about how well they would be able to understand Danish in new situations by condition (Subtitles + title, Subtitles no title, No subtitles + titles, No subtitles or title).



Note. Panel A displays subjects' mean ratings across the three key dependent variables measuring subject's ability to understand Danish in new situations (News, Weather, instructions). Panel B displays the pattern for each of the individual items comprising that mean and the "Friends" dependent variable. Error bars represent 95% confidence intervals of cell means.

Next, we looked for evidence to support our hypothesis that the title and subtitles boost confidence because they ease comprehension. We carried out 2(subtitles, no subtitles) x 2(title, no title) analyses on subjects' responses to "How difficult was it to comprehend the video?" (1 = *Very easy*, 7 = *Very difficult*). We found an interaction, such that titles only improved comprehension of the video when people had subtitles (F(1, 313) = 7.74, p = .01). Pairwise comparisons showed that compared to subjects who watched the Danish video with

no subtitles or title, subjects who watched the video with subtitles and a title reported the video was easier to comprehend (Tukey's HSD effect size: $M_{\text{diff}} = 2.55$ [1.94, 3.17], $p < .01$).

Finally, we calculated subjects' mean percentage score on the translation test to see if subjects who received both the title and subtitles learned any Danish. But their performance on the Danish translation test was no better than subjects in the other conditions (Danish translation test mean score: $M < 0.01\%$ in all conditions).

Experiment 4b

In Experiment 4b we addressed the question: To what extent would people who heard the Danish passage—knowing the topic and seeing English subtitles—mistake their ease of comprehension as evidence they could understand that language in new situations. We aimed to replicate the finding that subjects who watched the video with subtitles and a title reported they were better able to understand Danish more generally, compared to people who watched the Danish video with no subtitles or title, so in this experiment we only included those two conditions. We also removed all visual information from the video clips so subjects could hear only the foreign audio, isolating, and perhaps even emphasising, the effect of verbal context.

Method

Subjects. We used the Shiny Web app, "power for two independent groups t-test" (https://designingexperiments.com/shiny-r-web-apps/) to calculate sample size based on the data from Experiment 4a. We powered for what we thought would be the smallest effect size based on those data and a desired power of 0.90, the target *n* per condition was 144, or a total of 288. We collected data from 317 workers before exclusions. After exclusions, we analysed the data from 305 subjects (M_{age} = 40.00, SD_{age} = 13.74; 34% identified as men, 65% identified as women, 1% identified as gender diverse). 94% of subjects indicated that English was their first language.

Design. The experiment used a between-subjects design with two conditions (subtitles: title + subtitles, no title or subtitles).

Procedure. There were three phases in this experiment. In the first phase, subjects were randomly assigned to listen to one of three audio clips of people speaking a difficult to comprehend passage in Danish. Half of the subjects first saw the title of the passage (for example, "washing clothes") and then listened to the spoken Danish while English subtitles

appeared on the screen against a black background. The other half of subjects did not see the title, and only listened to the Danish audio. Each video depicted a black screen, but the subtitles group saw English subtitles in time with the audio clip.

In the second phase, subjects were asked the following three questions in a randomised order: "How well would you be able to follow Danish instructions in an emergency?" (*1 = Quite well, 5 = Not at all well*), "If you were listening to the Danish TV news, how well would you be able to understand the first story covered?", and "How well would you be able to understand the weather forecast on Danish TV?" Then we asked subjects to rate "If you were a new student at a Danish school and your peers spoke only Danish, how well would you be able to make friends?" (*1 = Quite well, 5 = Not at all well*). Because this "make friends" dependent variable did not replicate from Experiment 2 to Experiment 3, we speculated that asking this question in a random order might contaminate the other dependent variables, so it was always the fourth question.[6] Next, everyone was asked "How much expertise do you think is involved in understanding a foreign language?" (1 = *No expertise*, 5= *A great deal of expertise*) and "How difficult was it to comprehend the video?" (1 = *Very easy*, 5= *Very difficult*).

In the third phase, subjects listened to a series of Danish words (16 in total: 4 they had heard in the video, 6 from the videos they didn't watch, and 6 others that are commonly used) and were asked to translate these words into English.

Results and Discussion

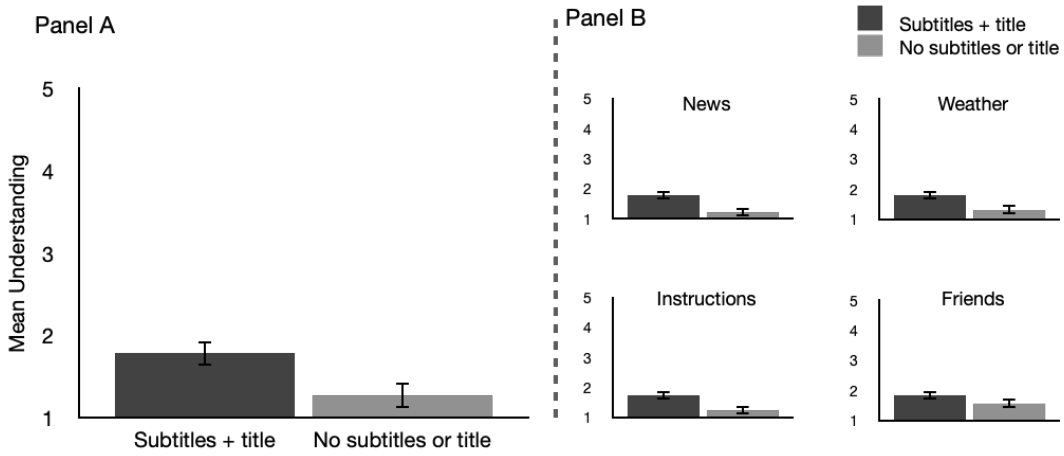
Recall our primary research question was: To what extent does watching a subtitled video of people speaking a foreign language inflate people's judgments about how well they can understand that language in new situations?

[6] In addition, a comment from an MTurk worker suggests that this item might be interpreted differently to the other dependent variables. "I have a comment: There have been several occasions in my life when I have made friends with people who did not speak English, and whose language I did not speak; we enjoyed each other's company, were able to communicate on a basic level, and taught other words in our native languages."

Before answering this question, we excluded from analysis data from 12 participants who [a] failed our compliance measures; or [b] reported speaking Danish. Across all exclusion criteria, 75% were excluded for failing one criterion and 25% for 2.

Figure 7

Subjects' mean ratings about how well they would be able to understand Danish in new situations by condition (title + subtitles, no title or subtitles).



Note. Panel A displays subjects' mean ratings across the three key dependent variables measuring subject's ability to understand Danish in new situations (News, Weather, instructions). Panel B displays the pattern for each of the individual items comprising that mean and the "Friends" dependent variable. Error bars represent 95% confidence intervals of cell means.

We now return to our primary research question. To answer this question, we first classified subjects' understanding ratings according to whether they saw the video with the title and subtitles or without the title and subtitles. We then calculated the mean across our three key understanding ratings and classified that mean according to whether subjects watched the subtitled video of people speaking Danish, or the video without subtitles, and display the findings in the main panel of Figure 7. As the figure shows, subjects who saw the

title and subtitles thought they would be better able to apply their newly-acquired comprehension of Danish to novel situations than those who saw neither the title nor subtitles, M_{diff} = 0.51, 95%CI [0.31, 0.70], Welch's $t(259) = 5.19$, $p < .001$; Wilcoxon $Z = 5.60$, $p < .001$.

As the Panel B of Figure 7 shows, we found that subjects who watched the video of people speaking Danish with a title and subtitles thought they would be better able to follow Danish instructions in an emergency, understand the first story on the Danish TV news, and understand the weather forecast on Danish TV than those who watched the video without a title or subtitles (all reverse coded: Instructions: M_{diff} = 0.49, 95%CI [0.27, 0.70], Welch's $t(259) = 4.49$, $p <.001$; Wilcoxon $Z = 4.78$, $p < .001$]; News: M_{diff} = .56, 95%CI [0.35, 0.77], Welch's $t(254) = 5.17$, $p <.001$ Wilcoxon $Z = 5.59$, $p < .001$; Weather: M_{diff} = .47, 95%CI [0.26, 0.69], Welch's $t(274) = 4.37$, $p <.001$; Wilcoxon $Z = 4.62$, $p < .001$). Subjects who saw a title and subtitles also thought they would be better able to make friends than subjects who saw neither (M_{diff} = 0.26, 95%CI [0.04, 0.48], Welch's $t(293) = 2.37$, $p = .02$; Wilcoxon $Z = 2.27$, $p = .02$). These findings provide further support for the idea that semantic context boosts people's confidence in their ability to apply their newly-acquired comprehension of Danish to novel situations.

In line with our prior data, we found support for the idea that the subtitles and title boosted people's comprehension of the video relative to people who didn't watch the video with subtitles and a title, (reverse coded: $M_{subtitles}$= 3.54, $M_{nosubtitles}$ =1.32, M_{diff} = 2.22, 95%CI [1.88, 2.56], Welch's $t(224) = 12.83$, $p < .001$; Wilcoxon $Z = 11.58$, $p <.001$). We also found people's ease of comprehension and mean subjective confidence ratings across the three key dependent variables were moderately associated ($r_{305} = -0.40$, 95% CI [−0.49, −0.30], $p < .001$).

But “subtitles + title” subjects did not perform better on the Danish translation test than their “no subtitles no title” counterparts. In other words, we found no evidence that people who watched the video with subtitles were learning any Danish from the short video. $M_{subtitles}$= 4.34%, $M_{nosubtitles}$ =5.42%, M_{diff} = -1.08%, 95%CI [-2.42, 0.26], Welch’s $t(302)$ = 1.59, p = .11; Wilcoxon Z = 1.79, p = .07.

Finally, regardless of whether or not subjects watched the video with a title and subtitles, subjects still rated the skill of understanding a foreign language as requiring considerable expertise, ($M_{subtitles}$ = 4.12, $M_{nosubtitles}$ = 3.89, M_{diff} = 0.22, 95%CI [-0.04, 0.49], Welch’s $t(294)$ = 1.64, p = .10; Wilcoxon Z = 1.30, p = .19), a finding consistent with our prior experiments.

Considered together, the results from Experiment 4b provide further support for the hypothesis that semantic context (operationalised here as a title and subtitles) enhances people’s comprehension of foreign language information in the moment. But people mistake this ease of comprehension as evidence they could better understand the foreign language more generally.

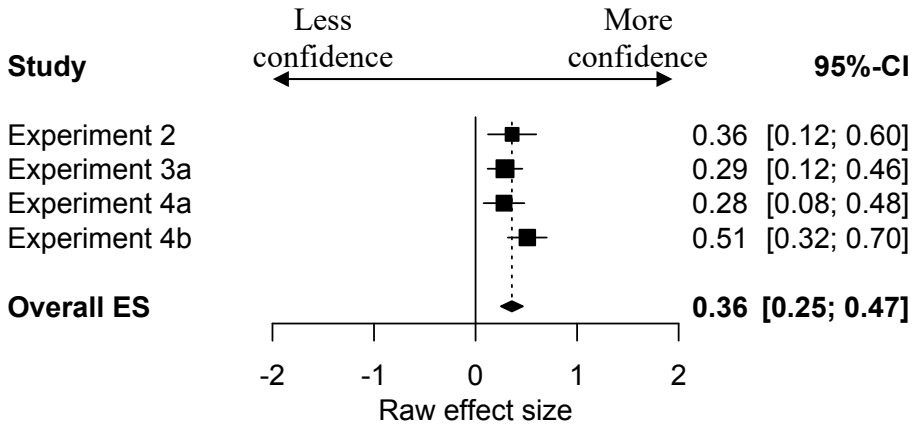
Chapter 3

Mini meta-analysis

To obtain a more precise estimate of the effect of watching a foreign video with subtitles on people's confidence in their ability to apply their newly-acquired comprehension of Danish to novel situations, we conducted a mini meta-analysis of the data from Experiments 2, 3a, 4a, and 4b (the experiments that measured subjective Danish understanding on a 5-point likert scale). We report the results in Figure 8.

Figure 8

Mini meta-analysis of subjects' mean understanding ratings by condition (subtitles, no subtitles) in Experiments 2, 3a, 4a, & 4b (R code; Carter & McCullough, 2014).



The right side of the vertical line in Figure 8 shows an effect in which subtitles (or in Experiment 4a and 4b, subtitles plus title) made subjects more confident in their ability to apply their newly-acquired comprehension of Danish to novel situations relative to receiving no subtitles (or in Experiment 4a and 4b, no subtitles or title). More specifically, we found a weighted raw effect size of $M_{\text{diff}} = .36$ [.25, .47], $p < 0.01$ or a 9% shift towards overconfidence. These findings fit with the idea that increased semantic context creates overconfidence in one's ability to understand a foreign language.

We can put this effect size in context by turning to the related literature. If we look to the cognitive literature, our effect size is similar to the increased false recall of words presented in a predictive semantic context (Whittlesea, 1993). Our effect sizes are also similar to people's illusion of better understanding complicated processes when they are presented with a picture (Cardwell, Lindsay & Garry, 2017). In the educational literature, our findings are in line with work showing students overconfidence in their understanding of scientific concepts accompanied by uninformative photographs (Wiley, 2019). Finally, our data are also in line with work showing people inflate their skill in dart-throwing, after watching 20 demonstrations of an expert throwing a dart at a bullseye (Kardas & O'Brien, 2018). Considered together, our findings align with the large body of work showing that fluent presentation can mislead people about how much they know or can do. And beyond the psychology literature, our effect size expressed as a correlation ($r = 0.215$) is similar to the relationship between conscientiousness scores and job proficiency, the effect of ibuprofen on pain reduction, and the likelihood of developing lung cancer among those who have ever smoked (see, for a review, Meyer et al., 2001).

Chapter 4

General Discussion

Across 7 experiments comprising 2387 subjects, we found evidence supporting the idea that increased semantic context can produce rapid illusions of foreign language ability. More specifically, when English speakers watched a Danish video with subtitles, they thought they would be better able to understand Danish in new situations, compared to people who watched the same video without subtitles.

In Experiment 1a and 1b, we found that watching a 1-minute foreign video-clip from a TV episode with English subtitles inflated people's estimates of how much dialogue they thought would understand if they were to watch another clip from that episode without subtitles.

In Experiment 2, we found subtitles produced a similar illusion of foreign language understanding when people were asked to evaluate how well they could apply that understanding in constrained novel situations related to the video, such as following directions and making friends.

In Experiment 3a we found this illusion persisted even when people evaluated how well they could apply their foreign language understanding in situations far removed from the content on the videos, such as understanding the Danish weather forecast or following Danish instructions. The findings from Experiment 3b support the idea that both subtitles, and the structure of the event portrayed in the video, might serve as forms of semantic context, easing people's comprehension of the video. These findings fit with the idea that the added semantic context provided by the video may enable subjects to more easily comprehend the video, but the misattribute this easier comprehension as evidence of understanding.

The findings from Experiment 4a and 4b support the idea that subtitles and titles provide semantic context that enhances people's comprehension of information in the moment. But people mistake this ease of comprehension as evidence they could understand the foreign language more generally. This conclusion fits with similar findings in the literature showing the dark side of semantic context (Laukkonen et al., 2020; Newman et al., 2012; Whittlesea, 1993). More specifically, they fit with the idea that semantic context can ease comprehension of information in the moment that people construe as evidence of learning or understanding (Cardwell, Newman, et al., 2017).

We also addressed several counterexplanations. For example, data from Experiment 3a suggest that this illusion is not produced because people actually learn Danish, nor because people thought they could default to English (rather than speak Danish) in the transfer situations. We also found the illusion did not hinge on a specific video. In fact, in Experiments 4a and 4b, we showed the illusion does not even hinge of video per se: we found similar effects when we minimized social information and eliminated sociogestural information. Considered together, then, our findings support the idea that semantic context produces rapid illusions of increased foreign language ability by making it easier for people to comprehend the situation in the moment—but then to confuse that for comprehending the foreign language more generally.

Implications

These findings contribute to several literatures. First, they extend what we know about the "dark side" of semantic context—how increased semantic context creates illusions of familiarity, truth, and understanding (Laukkonen et al., 2020; Newman et al., 2012; Whittlesea, 1993). In prior work, people were asked how well they understand how a specific, often complex, process works, such as how rainbows form (Cardwell, Lindsay & Garry, 2017). When an uninformative photo (say, a rainbow) accompanied the question,

people reported a better understanding of how that process works—even though subsequent analysis showed no such improved understanding. Here, we extend this work by showing that semantic context creates illusions not simply of familiarity or understanding, but also of rapidly-acquired skill, coupled with an ability to apply that skill to novel situations.

Second, our findings provide evidence of a potential mechanism for these sorts of illusions, in which semantic context produces feelings of easier comprehension which are then misattributed as evidence of understanding. More specifically, in Experiment 3b, people who watched a video clip with accompanying subtitles, or clearer event structure, reported the storyline was easier to follow and understand. Along the same lines, in Experiment 4b, people who watched subtitled audio (plus a title) of a spoken Danish passage reported easier comprehension. Collectively, our findings support the idea that people are influenced by increased semantic context—here, mistaking easier comprehension as evidence they would be better able to understand Danish in new situations, compared to others who did not have access to this semantic context.

Third, these findings also fit with work suggesting that processing fluency can be produced by a number of techniques, conceptual or perceptual, with predictable and similar effects on people's judgements (Alter & Oppenheimer, 2008; Stanley et al., 2021). Take judgements of truth, for example: people are more likely to say a statement is true when it rhymes; when it is paired with a photo; when words or concepts related to that statement are semantically primed, and when it is presented in high contrast (Kelley & Lindsay, 1993; Newman et al., 2012; McGlone & Tofighbakhsh, 2000; Reber & Schwarz, 1999). Likewise, processing fluency is also implicated in the well-known finding that people tend to incorrectly report that words presented on a computer appear for longer when those words are more frequent, primed, or repeated (for a review, see Fernandes & Garcia-Marques, 2020). Obviously, these manipulations do not give people "diagnostic" information about truth or

the duration, but the idea is that they all help to make it feel easier to process the information, and people misattribute that feeling to the task at hand (Unkelbach & Greifeneder, 2013). Here, subjects reported better ability to understand Danish in new situations after watching a video with accompanying subtitles (and sometimes a title), or a structure that made the storyline easier to follow. Moreover, this inflation in people's subjective Danish understanding ratings occurred in situations beyond the content they were viewing, and fits with the idea that people were making inferences regarding their own abilities more generally. And though we did not teach people Danish, our data suggest we made it easier to understand a video clip. The fact that perceptual and conceptual manipulations of fluency can produce similar outcomes suggests an avenue for future work—testing the possibility that perceptual manipulations, such as the clarity of the video, the contrast of the subtitles, or the quality of the audio track, would also change people's confidence about their Danish skill (Newman & Schwarz, 2018). Related work by Newman and Schwarz supports this possibility (2018): when people listened to "researchers" describing their work, those who heard a version with poor audio quality rated the work as less important, the researcher less intelligent, and the talk worse overall, compared to people who listened to otherwise identical versions with better audio quality. These findings suggest that perceptual manipulations might influence people's evaluations when they are presented with new information, such that features that ease processing are associated with more positive evaluations.

These findings also align with work from the metacognitive literature showing that judgements of learning can be influenced by easier processing. Both perceptual, and conceptual, manipulations similarly inflate people's assessments about what they will recall in the future. In one study, people listened to a series of words at varying volumes and made JOLs about each one (Rhodes & Castel, 2009). Those played at loud volumes were easier to process, and people awarded them higher JOLs, than words played at quiet volumes. But

people's recall of those words on a later test was no better. What's more, those JOLs seemed to reflect study behaviour; people were more likely to choose to restudy quiet words. Similar patterns arise when people are given word lists containing words of differing familiarity (Begg et al., 1989). People rate common words as easier to imagine and understand, and they award them higher JOLs than rare words. But their recall of common words is actually worse. Our findings suggest subtitles also produce easier comprehension of information that people misattribute as evidence of learning.

Our findings are reminiscent of the illusions of knowledge people demonstrate after searching Google for answers about specific topics. Later, when these same people are barred from using Google to answer questions about new topics, they overestimate how well they would be able to answer those new questions—a finding that suggests failure to recognize the influence of "outsourced" knowledge (Fisher et al., 2015). And this failure to distinguish one's own knowledge from external sources isn't limited to using the Internet: learning that scientists understand a given scientific phenomenon boosts people's appraisals of their own understanding of that phenomenon (Sloman & Rabb, 2016). Taken together, these findings suggest people conflate information out in the world with knowledge they possess. Likewise, our findings suggest that people provided with subtitles conflated that information with skills they possessed. These findings are all the more puzzling given the same people also reported that learning a foreign language requires a great deal of expertise. It seems plausible, then, that our subjects demonstrated "knowledge neglect," a failure to retrieve or apply relevant knowledge (Fazio et al., 2013; see Marsh & Umanath, 2014, for a review). More specifically, our findings suggest people fail to retrieve, or apply, their demonstrated knowledge of the expertise involved in understanding a foreign language, when evaluating their own ability to understand that language.

Of course, a critic might argue that subtitles rarely pushed people's estimates of their abilities into a level we would interpret as "high." Instead, what's important here is not people's estimates in absolute terms, but their reported increased ability after a single exposure to a short video. Moreover, people reported this increased ability despite also reporting knowledge that understanding a foreign language requires great deal of expertise. What then would be the effect of increasing exposure on confidence? We know from work on the misinformation effect that repeatedly hearing suggestions increases false memories of those suggestions (Zaragoza & Mitchell, 1996). In a similar vein, repeatedly imagining performing an action, such as breaking a toothpick, increases both people's likelihood of falsely reporting they actually performed that action and their confidence in that response (Goff and Roediger 1998, Thomas & Loftus, 2002). Likewise, work on the repetition truth effect suggests that repeating statements increases people's likelihood to rate those statements as true (Dechêne et al., 2010). Perhaps most relevant here, we know that repeatedly watching experts perform tasks—say, watching an expert dart-thrower 20 times—increases people's confidence in their own dart throwing skills compared with people who watch the video just once, even though they get no better (Kardas & O'Brien, 2018). It's plausible that with repeated watching—say, for example, watching an entire season of a foreign TV show—the illusion could compound. This possibility is an intriguing avenue for future research. Taken together, then, our findings have theoretical implications for our understanding of semantic context. processing fluency, and "knowledge neglect."

Our findings also have practical implications. Here, we provide evidence that ease of processing in the moment can distort people's metacognitive judgements about how well they understand information, which in turn changes their evaluations of how well they could apply that understanding in new situations. These findings also add to the literature showing that presenting the same information to people in different ways can change people's calibrations

of their own learning, performance, and abilities (Carpenter et al., 2013, Michael & Garry, 2019; Toftness et al., 2018; Weinstein & Roediger, 2012; Witherby & Carpenter, 2022). For instance, students draw on feelings of ease as evidence of learning—but those feelings of ease can inflate student's evaluations of their own skills— as a result, they invest less time and effort during learning (Dreisbach & Fischer, 2011; Luna & Albuquerque, 2022; Kornell, & Bjork, 2007; see, for a review, Chang & Brainerd, 2022). If semantic context eases processing of new information that people misattribute as evidence of their own skill, they may invest less time in their learning, potentially leading to poorer learning outcomes.

Although we addressed illusions of foreign language skill, it is reasonable to suspect similar effects might be revealed in other educational situations. For example, if students are learning about complex topics and technical terms in online lectures, captions might produce similar illusions of understanding. Therefore, educators might consider incorporating other features into their teaching—such as repeated testing and spaced/interleaved practice—to discourage overconfidence, help students recalibrate what they really do and do not understand, and even support the transfer of learning so people can apply that knowledge in new contexts including other test formats and knowledge domains (Carpenter, 2012; Carpenter & Olson, 2012; Kornell & Bjork, 2007, Roediger et al., 2011). "Desirable difficulties" such as these might make the learning process feel more difficult, but they have beneficial effects on long-term retention (Bjork & Bjork, 2011; Bjork & Bjork, 2020; Dunlosky et al., 2013). These desirable difficulties do not simply improve students' retention of fact-based knowledge either; they can also enhance their ability to analyse knowledge and apply it in new situations (Agarwal, 2019). In fact, recent work suggests that a detailed foundation of fact-based knowledge isn't essential for higher order learning (such as analysis or application) to take place; it can be enhanced by repeatedly testing students using more complicated application problems early in the learning process (Agarwal, 2019).

The effects we report here are new and surprising examples of rapidly-acquired illusions of skill. The data suggest that these illusions occur because providing enriching context, such as with subtitles, eases people's understanding of the videos—but people also misattribute this ease of understanding as evidence of their own abilities. Our work furthers our understanding about den mørke side af semantisk kontekst[7].

[7]the dark side of semantic context

References

Addis, D. R. (2020) Mental time travel? A neurocognitive model of event simulation. *Review of Philosophy and Psychology, 11*(2)**,** 233–259. https://doi.org/10.1007/s13164-020-00470-0

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189–209. https://doi.org/10.1037/edu0000282

Alicke, M. D., & Govorun, O. (2005). The Better-Than-Average Effect. In M. D. Alicke, D. A. Dunning, & J. I. Krueger (Eds.), *Studies in self and identity. The Self in Social Judgment* (pp. 85-106). New York, NY, US: Psychology Press.

Alter, A. L., & Oppenheimer, D. M. (2008). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*(3), 436–451. http://dx.doi.org/10.1037/a0020218

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718–735. https://doi.org/10.1037/a0029395

Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences*, *5*(12), 533–539. https://doi.org/10.1016/s1364-6613(00)01804-0

Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology*, *51*(5), 267.

Ball, B. H., Klein, K. N., & Brewer, G. A. (2014). Processing fluency mediates the influence of perceptual information on monitoring learning of educationally relevant

materials. *Journal of Experimental Psychology: Applied, 20*(4), 336–348. https://doi.org/10.1037/xap0000023

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, *116*(1), 261-292. https://doi.org/10.1162/003355301556400

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610-632. https://doi.org/10.1016/0749-596X(89)90016-8.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68. https://doi.org/10.1037/0096-3445.127.1.55

Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology, 59*(5), 960–970. https://doi.org/10.1037/0022-3514.59.5.960

Bisson, M. J., Van Heuven, W. J., Conklin, K., & Tunney, R. J. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics*, *35*(2), 399-418. https://doi.org/10.1017/S0142716412000434

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York: Worth Publishers.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475-479. https://doi.org/10.1016/j.jarmac.2020.09.003

Bø, S., & Wolff, K. (2020). I can see clearly now: Episodic future thinking and imaginability in perceptions of climate-related risk events. *Frontiers in Psychology*. *218*(11), 1–15. https://www.frontiersin.org/article/10.3389/fpsyg.2020.00218

Bower, G. H. (1978). Experiments on story comprehension and recall. *Discourse Processes*, *1*(3), 211-231. https://doi.org/10.1080/01638537809544437

Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema, 29*, 552-557. https://doi.org/10.7334/psicothema2016.383)

Bransford, J.D., & Johnson, M. (1972). Contextual prerequisites for understanding: Some Investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 717-726. https://doi.org/10.1016/S0022-5371(72)80006-9

Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, *194*(104054). https://doi.org/10.1016/j.cognition.2019.104054

Calvillo, D. P., Flores, A. N., Lara, P. M., Hawkins, W. C., Boman, L. D., & Smelter, T. J. (2019). Imagining experiencing an event in the future inflates certainty that it occurred in the past. *Imagination, Cognition and Personality, 39*(1), 5–24. https://doi.org/10.1177/0276236618803308

Cardwell, B. A., Lindsay, D. S., & Garry, M. (2017). Uninformative photos can increase people's perceived knowledge of complicated processes. *Journal of Applied Research in Memory and Cognition*, *6*(3), 244–252. https://doi.org/10.1016/j.jarmac.2017.05.002

Cardwell, B. A., Newman, E. J., Garry, M., Mantonakis, A., & Beckett, R. (2017). Photos that increase feelings of learning promote positive evaluations. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *43*(6), 944-954. https://doi.org/10.1037/xlm0000358

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279-283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 92–101. https://doi.org/10.1037/a0024828

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, *20*(6), 1350-1356. https://doi.org/10.3758/s13423-013-0442-z

Carter, E. C. & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology, 5*, Article 823. https://doi.org/10.3389/fpsyg.2014.00823

Castel, A.D., McCabe, D.P. & Roediger, H.L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review, 14*(1), 107–111. https://doi.org/10.3758/BF03194036

Chang, M., & Brainerd, C. J. (2022). Association and dissociation between judgments of learning and memory: A meta-analysis of the font size effect. *Metacognition and Learning*. Advance online publication. https://doi.org/10.1007/s11409-021-09287-3

Cross, K. P. (1977). Not can, but will college teaching be improved? *New Directions for Higher Education,* 1977(*17*), 1-15. https://doi.org/10.1002/he.36919771703

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*(2), 238–257. https://doi.org/10.1177/1088868309352251

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, *30*(1). https://doi.org/10.5334/irsp.82

Delivett, C. P., Klepacz, N. A., Farrow, C. V., Thomas, J. M., Raats, M. M., & Nash, R. A. (2020). Front-of-pack images can boost the perceived health benefits of dietary products. *Appetite*, *155*(104831). https://doi.org/10.1016/j.appet.2020.104831

Dreisbach, G., Fischer, R. (2011). If it's hard to read . . . try harder! Processing fluency as signal for effort adjustments. *Psychological Research*, *75*, 376–383. https://doi.org/10.1007/s00426-010-0319-y

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58. https://doi.org/10.1177/1529100612453266

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in Experimental Social Psychology,* Vol. 44, pp. 247–296). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83-87. https://doi.org/10.1111/1467-8721.01235

Eisenberg, M. L., Zacks, J. M., & Flores, S. (2018). Dynamic prediction during perception of everyday events. *Cognitive Research: Principles and Implications*, *3*(1), 1-12. https://doi.org/10.1186/s41235-018-0146-z

Engeler, I., & Häubl, G. (2021). Miscalibration in predicting one's performance: Disentangling misplacement and misestimation. *Journal of Personality and Social Psychology, 120*(4), 940–955. https://doi.org/10.1037/pspi0000301

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 540-551. https://doi.org/10.1016/S0022-5371(81)90165-1

Fagerland, M. W., & Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary Clinical Trials*, *30*(5), 490-496. https://doi.org/10.1016/j.cct.2009.06.007)

Fazio, L. K., Barber, S. J., Rajaram, S., Ornstein, P. A., & Marsh, E. J**.** (2013). Creating illusions of knowledge: Learning errors that contradict prior knowledge. *Journal of Experimental Psychology: General*, *142*(1), 1–5. https://doi.org/10.1037/a0028649

Federal Aviation Administration (FAA). (2016). *Airplane flying handbook* (FAA-H-8083-38). US Department of Transportation; Wickens, C.D, & Dehais, F. (2019). *Expertise in aviation*. Oxford, UK: Oxford University Press.

Fernandes, A. C., & Garcia-Marques, T. (2020). A meta-analytical review of the familiarity temporal effect: Testing assumptions of the attentional and the fluency-attributional accounts. *Psychological Bulletin*, *146*(3), 187-217. https://doi.org/10.1037/bul0000222

Fisher, M., Smiley, A. H., & Grillo, T. L. (2022). Information without knowledge: The effects of Internet search on learning. *Memory*, *30*(4), 375-387. https://doi.org/10.1080/09658211.2021.1882501

Gernsbacher, M. A. (2015). Video captions benefit everyone. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 195-202. https://doi.org/10.1177/2372732215602130

Gerrie, M. P., Belcher, L. E., & Garry, M. (2006). 'Mind the gap': False memories for missing aspects of an event. *Applied Cognitive Psychology, 20*(5), 689–696. https://doi.org/10.1002/acp.1221

Gerrig, R. J. (1993). *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press.

Goff, L. M., & Roediger, H. L. III. (1998). Imagination inflation for action events: Repeated imaginings lead to illusory recollections. *Memory & Cognition*, *26*(1), 20–33. https://doi.org/10.3758/BF03211367

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology*, *64*, 257-283. https://doi.org/10.1146/annurev-psych-113011-143802

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. J*ournal of Personality and Social Psychology, 79*(5), 701-721. https://doi.org/10.1037/0022-3514.79.5.701

Gregory, W. L., Cialdini, R. B., & Carpenter, K. M. (1982). Self-relevant scenarios as mediators of likelihood estimates and compliance: Does imagining make it so? *Journal of Personality and Social Psychology, 43*(1), 89-99. https://doi.org/10.1037/0022-3514.43.1.89

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 702–718. https://doi.org/10.1037/0278-7393.11.1-4.702

Hettmansperger, T.P. and McKean J.W. (2011). *Robust Nonparametric Statistical Methods*, 2nd ed., New York: Chapman-Hall.

Hirshleifer, D., Low, A., & Teoh, S. H. (2012). Are overconfident CEOs better innovators?. *The Journal of Finance*, *67*(4), 1457-1498. https://doi.org/10.1111/j.1540-6261.2012.01753.x

Hodges, B., Regehr, G., & Martin, D. (2001). Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine*, *76*(10 Suppl), S87–S89. https://doi.org/10.1097/00001888-200110001-00029

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*(3), 306–340. https://doi.org/10.1037/0096-3445.110.3.306

Johnson, D. D., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature, 477*, 317-320. https://doi.org/10.1038/nature10384

Johnson, D. D., Wrangham, R. W., & Rosen, S. P. (2002). Is military incompetence adaptive?: An empirical test with risk-taking behaviour in modern warfare. *Evolution and Human Behavior*, *23*(4), 245-264. https://doi.org/10.1016/S1090-5138(01)00087-3

Johnson, M. K., & Sherman, S. J. (1990). Constructing and reconstructing the past and the future in the present. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior,* Vol. 2, (pp. 482–526). The Guilford Press.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3-28. https://doi.org/10.1037/0033-2909.114.1.3.

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

Kardas, M., & O'Brien, E. (2018). Easier seen than done: Merely watching others perform can foster an illusion of skill acquisition. *Psychological Science*, *29*(4), 521–536. https://doi.org/10.1177/0956797617740646

Kellerman, S. (1992). 'I see what you mean': The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied linguistics*, *13*(3), 239-258. https://doi.org/10.1093/applin/13.3.239

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*(1), 1–24. https://doi.org/10.1006/jmla.1993.1001

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 483-502). New York: Guilford Publications.

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1133–1145. https://doi.org/10.1037/0278-7393.32.5.1133

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. https://doi.org/10.3758/bf03194055

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*(6), 787–794. https://doi.org/10.1177/0956797611407929

Krawczyk, M., & Wilamowski, M. (2017). Are we all overconfident in the long run? Evidence from one million marathon participants. *Journal of Behavioral Decision Making*, *30*(3), 719-730. https://doi.org/10.1002/bdm.1984

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Lackner, M., & Sonnabend, H. (2020). Gender differences in overconfidence and decision-making in high-stakes competitions: Evidence from freediving contests. *ResearchGate*. https://doi.org/10.13140/RG.2.2.34496.97284

Lakshmanan, A., & Krishnan, H. S. (2011). The Aha! Experience: Insight and Discontinuous Learning in Product Usage. *Journal of Marketing*, *75*(6), 105–123. https://doi.org/10.1509/jm.10.0348

Laukkonen, R. E., Kaveladze, B. T., Tangen, J. M., & Schooler, J. W. (2020). The dark side of Eureka: Artificially induced aha moments make facts feel true. *Cognition, 196,* Article 104122. https://doi.org/10.1016/j.cognition.2019.104122

Lindsay, D. S. (2014). Memory source monitoring applied. In Perfect, T. J., & Lindsay, D. S. (Eds.). *The SAGE handbook of applied memory*, (pp. 59-75). Sage.

Lindsay, D.S. (2008). Source Monitoring. In J.H. Byrne (Ed.), *Learning and memory: A comprehensive reference*, (pp. 325-347). Elsevier, Oxford: UK. https://doi.org/10.1016/B978-012370509-9.00175-3.

Luna, K & Albuquerque, P. (2022). Do beliefs about font size affect retrospective metamemory judgments in addition to prospective judgments?: A registered report. *Experimental Psychology.* Advance online publication. https://doi.org/10.1027/1618-3169/a000549.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology, 88*(1), 49–63. https://doi.org/10.1037/0022-0663.88.1.49

Marsh, E. J., & Umanath, S. (2014). Knowledge neglect: Failures to notice contradictions with stored knowledge. In D. N. Rapp and J. Braasch (Eds) *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences* (p. 161-180). MIT Press.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1-23. https://doi.org/10.3758/s13428-011-0124-6

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology, 82*(4), 715-726. https://doi.org/10.1037/0022-0663.82.4.715

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. *Psychological Science*, *11*(5), 424-428. https://doi.org/10.1111/1467-9280.00282

McKay, R. T., & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences, 32*(6), 493-510. https://doi.org/10.1017/S0140525X09990975.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., ... & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist, 56*(2), 128.

Michael, R. B., & Garry, M. (2019). How do ordered questions bias eyewitnesses? *Memory*, *27*(7), 904–915. https://doi.org/10.1080/09658211.2019.1607388

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect."

Psychological Science, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Newman, E.J., Azad, T., Lindsay, D.S., & Garry, M. (2018). Evidence that photos promote rosiness for claims about the future. *Memory & Cognition*, *46*(8), 1223–1233. https://doi.org/10.3758/s13421-016-0652-5

Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, *19*(5), 969-974. https://doi.org/10.3758/s13423-012-0292-0

Newman, E. J., Garry, M., Unkelbach, C., Bernstein, D. M., Stephen Lindsay, D., & Nash, R. A. (2015). Truthiness and falsiness of trivia claims depend on judgmental contexts. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(5), 1337–1348. https://doi.org/10.1037/xlm0000099

Newman, E. J., Jalbert, M. C., Schwarz, N., & Ly, D. P. (2020). Truthiness, the illusory truth effect, and the role of need for cognition. *Consciousness and Cognition: An International Journal, 78,* Article 102866. https://doi.org/10.1016/j.concog.2019.102866

Newman, E. J., & Schwarz, N. (2018). Good sound, good research: How audio quality influences perceptions of the research and researcher. *Science Communication*, *40*(2), 246-257. https://doi.org/10.1177/1075547018759345

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237–241. https://doi.org/10.1016/J.TICS.2008.02.014

Paulhus, D. L. (2011). Overclaiming on personality questionnaires. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.). *New perspectives on faking in personality assessment* (pp. 151–164). New York: Oxford University Press.

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence*, *32*(3), 297–314. https://doi.org/10.1016/j.intell.2004.02.001

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890–904. https://doi.org/10.1037/0022-3514.84.4.890

Perego, E., Del Missier, F., Porta, M., & Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media Psychology*, *13*(3), 243-272, https://doi.org/10.1080/15213269.2010.502873

Persky, A. M., Lee, E., & Schlesselman, L. S. (2020). Perception of learning versus performance as outcome measures of educational research. *American Journal of Pharmaceutical Education*, *84*(7), ajpe7782. https://doi.org/10.5688/ajpe7782

Pichert, J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. *Journal of Educational Psychology*, *69*(4), 309-315. https://doi.org/10.1037/0022-0663.69.4.309

Price, A., Lindholm, T., Gjervig Gram, J. (Writers) & Kragh-Jacobsen, S. (Director). (2010) Season 1, Episode 1. [Television series episode] In Hammerich, C. (Producer), *Borgen*. Denmark: DR Fiktion.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*, 338-342. https://doi.org/10.1006/ccog.1999.0386

Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory and Language, 30*(4), 385–406. https://doi.org/10.1016/0749-596X(91)90013-A

Rhodes, M. G. & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615–625. https://doi.org/10.1037/a0013684

Rhodes, M.G., Castel, A.D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*, 550–554. https://doi.org/10.3758/PBR.16.3.550

Roediger, H. L. III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Royer, J. M. (1986). Designing instruction to produce understanding: An approach based on cognitive theory. In G. D. Phye & T. Andre (Eds.), *Cognitive classroom learning: Understanding, thinking, and problem solving* (pp. 83–113). Orlando, FL: Academic Press.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521-562. https://doi.org/10.1207/s15516709cog2605_1

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10-28. http://dx.doi.org/10.1037/pspa0000102

Sandro Rota. (2017, May 10). *Embraer E190. HC-CGG. Catamayo / Loja (SECA) - Quito (SEQM)* [Video]. Youtube. https://www.youtube.com/watch?v=z1Dyy2qUdro

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the*

Royal Society B: Biological Sciences, *362*(1481), 773-786. https://doi.org/10.1098/rstb.2007.2087

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, *8*(9), 657-661. https://doi.org/10.1038/nrn2213.

Schunk, D. H. (2012). *Learning theories an educational perspective sixth edition*. Pearson.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist, 54*, 93-105. https://doi.org/10.1037/0003-066X.54.2.93

Schwarz, N. (2012). Feelings-as-information theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 289–308). Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n15

Schwarz, N. (2018). of fluency, beauty, and truth: Inferences from metacognitive experiences. In J. Proust M. Fortier (eds.), *Metacognitive Diversity: An interdisciplinary approach* (pp. 25 - 46). Oxford: Oxford University Press.

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 1074–1083. https://doi.org/10.1037/0278-7393.18.5.1074

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy, 2*(1), 85-95. https://doi.org/10.1353/bsp.2016.0009.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18*(7), 698-711. https://doi.org/10.1080/09658211.2010.506441

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

Sherman, S. J., Cialdini, R. B., Schwartzman, D. F., & Reynolds, K. D. (1985). Imagining can heighten or lower the perceived likelihood of contracting a disease: The mediating effect of ease of imagery. *Personality and Social Psychology Bulletin, 11*(1), 118–127. https://doi.org/10.1177/0146167285111011

Simons, D. J. (2013). Unskilled and optimistic. *Psychonomic Bulletin Review*, *20*(3), 601-607. https://doi.org/10.3758/s13423-013-0379-2

Sloman, S. A., & Rabb, N. (2016). Your understanding is my understanding: Evidence for a community of knowledge. *Psychological Science*, *27*(11), 1451-1460. https://doi.org/10.1177/0956797616662271

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288-307. https://doi.org/10.1006/obhd.2000.2926

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Song, H., & Schwarz, N. (2008). If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological Science, 19*(10), 986–988. https://doi.org/10.1111/j.1467-9280.2008.02189.x

Stanley, M. L., Whitehead, P. S., & Marsh, E. J. (2022). The cognitive processes underlying false beliefs. *Journal of Consumer Psychology*, *32*(2), 359-369. https://doi.org/10.1002/jcpy.1289

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel and is it unique to humans? *Behavioral and Brain Sciences, 30*(3)*,* 299–313. https://doi.org/10.1017/S0140525X07001975

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661-699. Doi:10.1111/j.0023-8333.2005.00320.x

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193-210. https://doi.org/10.1037/0033-2909.103.2.193

Thomas, A. K., & Loftus, E. F. (2002). Creating bizarre false memories through imagination. *Memory & Cognition*, *30*(3), 423–431. https://doi.org/10.3758/BF03194942

Toftness, A. R., Carpenter, S. K., Geller, J., Lauber, S., Johnson, M., & Armstrong, P. I. (2018). Instructor fluency leads to higher confidence in learning, but not better learning. *Metacognition and Learning*, *13*(1), 1-14. https://doi.org/10.1007/s11409-017-9175-0

Torpe, C. (Writer), Kaalund, L. (Director). (2012) Season 1, Episode 1. [Television series episode] In Leth, K., & Morthorst, J. (Producers), *Rita*. Denmark: SF Studios.

Townsend, C. L., & Heit, E. (2011). Judgments of learning and improvement. *Memory & Cognition*, *39*(2), 204–216. https://doi.org/10.3758/s13421-010-0019-2

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne, 26*(1), 1–12. https://doi.org/10.1037/h0080017

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

Unkelbach, C. (2006). The learned interpretation of cognitive fluency. *Psychological Science, 17*(4)*,* 339-345. https://doi.org/10.1111/j.1467-9280.2006.01708.x

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(1), 219–230. https://doi.org/10.1037/0278-7393.33.1.219.

Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In. C.Unkelbach, & R. Greifeneder. *The experience of thinking: How the fluency of mental processes influences cognition and behaviour*, (pp. 11-32). London: Psychology Press.

Weinstein, Y., & Roediger, H. L. (2010). Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition*, *38*(3), 366–376. https://doi.org/10.3758/MC.38.3.366

Weinstein, Y., & Roediger, H. L. (2012). The effect of question order on evaluations of test performance: how does the bias evolve? *Memory & Cognition, 40*(5), 727-735. https://doi.org/10.3758/s13421-012-0187-3

Whittlesea, B. W. A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*(6)*,* 1235–1253. https://doi.org/10.1037/0278-7393.19.6.1235

Wiley, J. (2019). Picture this! Effects of photographs, diagrams, animations, and sketching on learning and beliefs about learning from a geoscience text. *Applied Cognitive Psychology*, *33*, 9-19. https://doi.org/10.1002/acp.3495

Wiley, J., & Rayner, K. (2000). Effects of titles on the processing of text and lexically ambiguous words: Evidence from eye movements. *Memory & Cognition*, *28*(6), 1011–1021. https://doi.org/10.3758/BF03209349

Williams, A. F. (2003). Views of US drivers about driving safety. *Journal of Safety Research, 34*(5), 491-494. https://doi.org/10.1016/j.jsr.2003.05.002

Wilson, J. C., & Westerman, D. L. (2018). Picture (im)perfect: Illusions of recognition memory produced by photographs at test. *Memory & Cognition*, *46*(7), 1210–1221. https://doi.org/10.3758/s13421-018-0832-6

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (eds.), *The psychology of evaluation: Affective processes in cognition and emotion*, pp. 189–217). Lawrence Erlbaum Associates Publishers.

Witherby, A. E., & Carpenter, S. K. (2022). The impact of lecture fluency and technology fluency on students' online learning and evaluations of instructors. *Journal of Applied Research in Memory and Cognition*. Advance online publication. https://doi.org/10.1037/mac0000003

YouGov. (2019) Do you think if you were playing your very best tennis, you could win a point off Serena Williams. [Data set]. Retrieved from: https://yougov.co.uk/topics/lifestyle/survey-results/daily/2019/07/12/ce84b/1

YouGov. (2021). Humans vs Animals [Data set]. Retrieved from: https://yougov.co.uk/topics/lifestyle/articles-reports/2021/05/21/which-animals-could-britons-beat-fight.

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—And is not—A desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition, 41*(2), 229–241. https://doi.org/10.3758/s13421-012-0255-8

Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, *71*, 165-191. https://doi.org/10.1146/annurev-psych-010419-051101

Zenger, T. R. (1992). Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover. *Administrative Science Quarterly, 37*(2), 198–219. https://doi.org/10.2307/2393221

Appendix A

Table 1
Additional information about videos used in Part 1 and Part 2.

Video	Synopsis	Length (min:sec)	Source
Plane landing	Cut down cockpit footage of plane landing. Limited view of pilot.	3:44	Sandro Rota Youtube channel (Sandro Rota, 2017)
Rita Danish video	A teacher shows a new hire around a school, and everyone is speaking Danish.	1:00	Television show *Rita* (Torpe & Kaalund, 2012)
Borgen Danish Video	Several politicians engaged in a live, chaotic, televised, debate.	1:00	Television show *Borgen* (Price et al., 2010)

Appendix B

Preregistration for Part 1, Experiment 2

'Confidence in ability to perform highly skilled tasks '
(AsPredicted #23118)

Created: 05/06/2019 06:58PM(PT)
Made Public: 03/02/2022 01:46 PM (PT)
https://aspredicted.org/z2az3.pdf

Author(s)
Kayla Jordan (University of Waikato) - krj13@students.waikato.ac.nz

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does watching a pilot land a plane inflate people's confidence in their own ability to land a plane?

3) Describe the key dependent variable(s) specifying how they will be measured.
We will measure subjects' confidence in their ability to land the plane in two ways:
1) "How confident are you that you could land the plane without dying?" (0% = not at all confident, 100% = very confident)
2) "How confident are you that you could land the plane as well as a pilot could?" (0% = not at all confident, 100% = very confident)
We will counterbalance the order of these two questions.

4) How many and which conditions will participants be assigned to?
The study will use a between subjects design with two conditions (Video type:Video, no video).

Subjects will either watch a video of a pilot landing a plane before making their confidence ratings or they will watch no such video.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will conduct a between subjects independent t-test, and post-hoc tests, on subjects' confidence ratings with Video type(video, no video) as a between subjects factor.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude from analysis the data from subjects who:
A) provide a nonsensical description of the situation they were asked to imagine
B) have flown or landed a plane before
C) have a pilot's license

D) fail the attention check we have built into our ratings questionnaire

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.
Per the Shiny Web sample size for power for two independent groups t-test app (https://designingexperiments.com/shiny-r-web-apps/): based on a previous study of ours with a between-subjects design with two groups, Mdiff=8.59, common SD= 29, with a desired power 0.9, the required n per group = 226. Therefore, factoring in anticipated exclusions, we will collect data from 600 participants total.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We are collecting within subjects data on subjects' confidence ratings before and after watching the video, and will use those data for exploratory follow-up analyses.

We will conduct a between subjects two-way Anova, and post-hoc tests, on subjects' confidence ratings with Video (video, no video) and Gender as between subjects measures.

We will also ask subjects how easy it was for them to imagine landing the plane.

Appendix C

Preregistration for Part 2, Experiment 1a
'Subtitles experiment 1'
(AsPredicted #27600)

Created: 09/08/2019 04:23 PM (PT)

Author(s)
Kayla Jordan (University of Waikato) - krj13@students.waikato.ac.nz

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does watching a video with subtitles of people speaking a foreign language inflate people's judgments about how well they can understand that language?

3) Describe the key dependent variable(s) specifying how they will be measured.
We will measure subjects' language ability judgements in two ways:
1) "Later on in this survey we are going to show you another short video clip, without subtitles, of people speaking Danish. This new video clip is from the same episode of this TV show. How much of the spoken dialogue do you think you will understand in this video?" (0% = none of the dialogue 100% = all of the dialogue)
2) "Later on in this survey we are going to show you another short video clip, without subtitles, of people speaking Danish. How much of the spoken dialogue do you think you will understand in this clip?" (0% = none of the dialogue 100% = all of the dialogue)
Subjects will be randomly assigned to answer one of the 2 questions.

4) How many and which conditions will participants be assigned to?
The study will use a between subjects design with two conditions (Subtitles: subtitles, no subtitles).

Subjects will either watch a video with subtitles of people speaking a foreign language before making their language ability rating or they will watch the video without subtitles.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will conduct a between subjects independent t-test, on subjects' language ability ratings with Subtitles(subtitles, no subtitles) as a between subjects factor.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude from analysis the data from subjects who:
Provide a nonsensical description of the situation shown in the video
Speak the foreign language
Fail the attention check we have built into our ratings questionnaire

Have watched the series "rita" before

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about exactly how the number will be determined.
We will collect data from 400 ps, 100 per cell.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will conduct a between subjects two-way Anova, and post-hoc tests, on subjects' language ability ratings with Subtitles(subtitles, no subtitles) and Gender as between subjects measures.
We will also ask subjects how much they learnt from watching the video.

Appendix D

Preregistration for Part 2, Experiment 1b
'Subtitles experiment 1 replication'
(AsPredicted #81246)

Created: 11/28/2021 03:12 PM (PT)

Author(s)
Kayla Jordan (The University of Waikato) - krj13@students.waikato.ac.nz
Maryanne Garry (The University of Waikato) - maryanne.garry@waikato.ac.nz

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does watching a Danish video with subtitles inflate people's judgments about how well they can understand that language?

3) Describe the key dependent variable(s) specifying how they will be measured.
Later on in this survey we are going to show you another short video clip, without subtitles, of people speaking Danish. This new video clip is from the same episode of this TV show. How much of the spoken dialogue do you think you will understand in this video?" (0% = none of the dialogue 100% = all of the dialogue)

4) How many and which conditions will participants be assigned to?
The study will use a between subjects design with two conditions (Subtitles: subtitles, no subtitles).

Subjects will either watch a video with subtitles of people speaking a foreign language before making their language ability rating or they will watch the video without subtitles.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will conduct a between-subjects independent t-test, on subjects' language ability rating with Subtitles(subtitles, no subtitles) as a between-subjects factor.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude from analysis the data from subjects who:
Provide a nonsensical description of the situation shown in the video
Speak the foreign language
Fail the attention check we have built into our ratings questionnaire
Have watched the series "Rita" before

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.
Per the Shiny Web sample size for power for two independent groups t-test app

(https://designingexperiments.com/shiny-r-web-apps/): based on a between-subjects design with two groups, Mdiff=4, common SD= 15.78, with a desired power 0.95, the required n per group = 405. Therefore, factoring in anticipated exclusions, we will collect data from 850 participants total.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

Appendix E

Preregistration for Part 2, Experiment 2
'Subtitles experiment 2. '
(AsPredicted #31011)

Created: 11/12/2019 02:58 PM (PT)

Author(s)
Kayla Jordan (University of Waikato) - krj13@students.waikato.ac.nz

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does watching a video with subtitles of people speaking a foreign language inflate people's judgments about how well they can understand that language?

3) Describe the key dependent variable(s) specifying how they will be measured.
We will measure subjects' language ability judgements in two ways:
"Now insert yourself into this scenario and imagine interacting with these people."
1) "If one of these teachers were to give you directions to a classroom, how well would you be able to follow those directions?" (1 = Quite well, 5 = Not at all well)
2) "If you were a new student at this school, how well would you be able to make friends?" (1 = Quite well, 5 = Not at all well)
The order of these questions will be counterbalanced

4) How many and which conditions will participants be assigned to?
The study will use a between subjects design with two conditions (Subtitles: subtitles, no subtitles).

Subjects will either watch a video with subtitles of people speaking a foreign language before making their language ability ratings or they will watch the video without subtitles.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will conduct a between subjects independent t-test, on subjects' language ability ratings with Subtitles(subtitles, no subtitles) as a between subjects factor.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude from analysis the data from subjects who:
provide a nonsensical description of the situation shown in the video
Speak the foreign language
fail the attention check we have built into our ratings questionnaire

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about exactly how the number will be determined.
We will collect data from 300 subjects.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will conduct a between subjects two-way Anova, and post-hoc tests, on subjects' language ability ratings with Subtitles(subtitles, no subtitles) and Gender as between subjects measures.
We will also ask subjects how much they learnt from watching the video.

Appendix F

Preregistration for Part 2, Experiment 3a
'Subtitles danish test'
(AsPredicted #42085)

Created: 06/01/2020 01:49 AM (PT)

Author(s)
Kayla Jordan (University of Waikato) - krj13@students.waikato.ac.nz

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does watching a video with subtitles of people speaking a foreign language inflate people's judgments about how well they can understand that language? Do people learn more Danish after watching a 1minute Danish clip with subtitles compared to people who watch that clip without subtitles?

3) Describe the key dependent variable(s) specifying how they will be measured.
We will measure subjects' language ability judgements in four ways: 1) "How well would you be able to follow Danish instructions in an emergency? " (1 = Quite well, 5 = Not at all well) 2) "If you were a new student at a Danish school and your peers spoke only Danish, how well would you be able to make friends?" (1 = Quite well, 5 = Not at all well) 3) "How well would you be able to understand the top story on the Danish TV news?" 4) "How well would you be able to understand the weather forecast on Danish TV?" The order of these questions will be counterbalanced We will also measure people's understanding of Danish words with a short test at the end of the survey containing both words or short phrases from the 1-minute clip subjects have watched and commonly used words in the Danish language.

4) How many and which conditions will participants be assigned to?
The study will use a between-subjects design with two conditions (Subtitles: subtitles, no subtitles). Subjects will either watch one of two videos with subtitles of people speaking a foreign language before making their language ability ratings or they will watch one of the two videos without subtitles.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will conduct a between-subjects independent t-test, on subjects' language ability ratings with Subtitles(subtitles, no subtitles) as a between-subjects factor. We will also conduct a between-subjects independent t-test, on the mean score across the four language ability ratings with Subtitles(subtitles, no subtitles) as a between-subjects factor. We will sum up subjects' scores on the Danish language test and conduct a between-subjects independent t-test, on these scores with Subtitles(subtitles, no subtitles) as a between-subjects factor.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude from analysis the data from subjects who: Fail to complete the study provide

a nonsensical description of the situation shown in the video Speak the foreign language Watched the Danish television series “Rita” or “Borgen”

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

Per the Shiny Web sample size for power for two independent groups t-test app (https://designingexperiments.com/shiny-r-web-apps/): based on a previous study of ours with a between-subjects design with two groups, Mdiff=0.48, common SD= 1.31, with a desired power 0.9, the required n per group = 157. Therefore, factoring in anticipated exclusions, we will collect data from 350 participants total.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

We will conduct a between-subjects two-way Anova, and post-hoc tests, on subjects’ language ability ratings with Subtitles(subtitles, no subtitles) and Gender as between-subjects measures. We will also ask subjects how much they learnt from watching the video.

Appendix G

Preregistration for Part 2, Experiment 4a
'The effect of increased semantic context on language ability judgements'
(AsPredicted #53593)

Created: 12/06/2020 06:00 PM (PT)

Author(s)
Kayla Jordan (University of Waikato) - krj13@students.waikato.ac.nz

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does watching a danish video with subtitles and a picture and title that provide semantic context inflate people's judgments about how well they can understand that language?
Do people learn more Danish after watching a 1minute Danish clip with subtitles compared to people who watch that clip without subtitles?

3) Describe the key dependent variable(s) specifying how they will be measured.
We will measure subjects' language ability judgements in four ways:
1) "How well would you be able to follow Danish instructions in an emergency? " (1 = Quite well, 5 = Not at all well)
2) "If you were a new student at a Danish school and your peers spoke only Danish, how well would you be able to make friends?" (1 = Quite well, 5 = Not at all well)
3) "How well would you be able to understand the top story on the Danish TV news?"
4) "How well would you be able to understand the weather forecast on Danish TV?"
The order of these questions will be counterbalanced
We will also measure people's understanding of Danish words with a short test at the end of the survey containing both words or short phrases from the 1 minute clip subjects have watched and commonly used words in the Danish language.

4) How many and which conditions will participants be assigned to?
The study will use a 2x2 between subjects design (Subtitles: subtitles, no subtitles, Picture+title: picture and title, no picture or title).

Subjects will listen to one of 3 difficult to comprehend passages in Danish. These audio clips will be presented either with English subtitles and a picture and title, with subtitles but no picture or title, with a picture and title but without subtitles, or without a picture and title and without subtitles.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will conduct a between subjects ANOVA, and post-hoc tests, on each one of subjects' language ability ratings with Subtitles(subtitles, no subtitles) and Context(title and picture, no title or picture) as between subjects factors.
We will also conduct a between subjects ANOVA, and post-hoc tests, on the mean score

across the four language ability ratings with Subtitles(subtitles, no subtitles) and Context(title and picture, no title or picture) as between subjects factors.
We will sum up subjects' scores on the Danish language test and conduct a between subjects ANOVA, and post-hoc tests, on these scores with Subtitles(subtitles, no subtitles) and Context(title and picture, no title or picture) as between subjects factors.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude from analysis the data from subjects who:
Fail to complete the study
Fail the attention check built into our questionnaire
Provide a nonsensical description of the passage they listened to
Speak the foreign language

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.
Per the Gpower app for an Anova: based on a previous study of ours with a 2x2 between-subjects design, effect size f=0.22, number of groups=4, with a desired power 0.9, the required total sample size was 302. Therefore, factoring in anticipated exclusions, we will collect data from 315 participants total.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will conduct a between subjects two-way Anova, and post-hoc tests, on subjects' language ability ratings with Subtitles(subtitles, no subtitles) and Gender as between subjects measures.
We will also ask subjects how much they learnt from watching the video.
We will run correlations between subject's passage understanding ratings and language ability judgements(each one individually and mean across all 4)

Appendix H

Preregistration for Part 2, Experiment 4b
'Subtitles experiment-manipulating semantic context (subtitles and title, no subtitles or title).' (AsPredicted #86198)

Created: 01/25/2022 02:29 PM (PT)

Author(s)
Kayla Jordan (The University of Waikato) - krj13@students.waikato.ac.nz
Maryanne Garry (The University of Waikato) - maryanne.garry@waikato.ac.nz

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does watching a Danish video with subtitles and a title that provides semantic context inflate people's judgments about how well they can understand that language?
Do people learn more Danish after watching a 1minute Danish clip with subtitles compared to people who watch that clip without subtitles?

3) Describe the key dependent variable(s) specifying how they will be measured.
We will measure subjects' language ability judgements in three ways:
1) "How well would you be able to follow Danish instructions in an emergency? " (1 = Quite well, 5 = Not at all well)
2) "How well would you be able to understand the top story on the Danish TV news?"
3) "How well would you be able to understand the weather forecast on Danish TV?"
The order of these questions will be counterbalanced
We will also measure people's understanding of Danish words with a short test at the end of the survey containing both words or short phrases from the 1 minute clip subjects have watched and commonly used words in the Danish language.

4) How many and which conditions will participants be assigned to?
The study will use a between subjects design with 2 conditions (Context: subtitles and title, no subtitles or title).

Subjects will listen to one of 3 difficult to comprehend passages in Danish. These audio clips will be presented either with English subtitles and a title or without a title and without subtitles.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will conduct a between subjects t-test, on each one of subjects' language ability ratings with Context(subtitles and title, no subtitles or title) as a between subjects factor.
We will also conduct a between subjects t-test, on the mean score across the three language ability ratings with Context(subtitles and title, no subtitles or title) as a between subjects factor.
We will sum up subjects' scores on the Danish language test and conduct a between subjects t-test, on these scores with Context(subtitles and title, no subtitles or title) as a between

subjects factor.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude from analysis the data from subjects who:
Fail to complete the study
Fail the attention checks
Speak the foreign language

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.
Per the Shiny Web sample size for power for two independent groups t-test app (https://designingexperiments.com/shiny-r-web-apps/): we powered for what we thought would be the smallest effect size based on a previous study of ours with a between-subjects design with two groups, Mdiff=.38, common SD= .89, with a desired power 0.95, the required n per group = 144. Therefore, factoring in anticipated exclusions, we will collect data from 300 participants total.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
After the 3 main dependent variables listed above, we will also ask subjects: "If you were a new student at a Danish school and your peers spoke only Danish, how well would you be able to make friends?" (1 = Quite well, 5 = Not at all well)

Because this "make friends" item showed different patterns and was less correlated with the other dependent measures we put it last–reasonable balance of wanting to determine if the atypicality of the item would replicate here, but reducing it's impact on the other items.

Appendix I

Transcript of "Rita" video clip used in Experiment 1a, 1b, 2 and 3.

Danish exact speak	**Current English subtitles**
Godmorgen	
Godmorgen	Morning.
Godmorgen, Rita	Good morning, Rita.
Godmorgen	Good morning.
Godmorgen	
Det her, det er Hjørdis, som jeg fortalte dig om	Remember I told you about Hjørdis?
Hej	Hi.
Hjørdis, hej	
Hjørdis er frisk fra seminariet og skal være en del af lærergruppen i din fjerde- og niende klasse.	Straight out of teacher's college. She'll be part of the team in your 4th and 9th grade.
Så hvis du gider og vise hende lidt rundt – introducere hende i klassen	Would you show her around? Introduce her?
Ja, selvfølgelig	Of course.
Godt	Right.
Godmorgen	Morning.
Godmorgen, jeg kommer nu	Morning. I'll be right there.
Jamen, så må I have rigtig god fornøjelse	Have a nice day.
Øh – Lisbeth?	Lisbeth…
Ja, jamen øh. Har Rasmus fortalt dig, hvordan det fungerer her?	Did Rasmus fill you in.
Ja, det hele	Yes, I think so.
Glemt alt, hvad han har sagt	Forget what he said.
Han er sød, men han har aldrig undervist	He's nice, but he's never been a teacher.
Kommer du med?	Are you coming?
Hej Rita	Hi, Rita.
Hej Albert. Opfør dig ordentligt	Hi, Albert. Behave yourself.
Ja, det skal jeg nok.	Yes, I will.
Dernede, der har vi skolelederens kontor, og det er Rasmus, ham har du mødt.	That's the principal's office. Rasmus and his secretary.
Han sidder dernede med sin sekretær	
Og her har vi så aulaen, hvor vi holder fællesmøder	This is the common area where we have assemblies.
Skoleboden. Elevavisen holder til derovre. Computerne – der sidder de mærkelige børn.	The café, the computer where the weird kids hang out.
Hej Jeppe.	Hi, Jeppe.

Appendix J

Transcript of "Borgen" video used in Experiments 3a and 3b

Danish exact speak	Current English subtitles
Så er vi kommet til det, venner. Det er nu, alle partilederne får deres 2 minutters fri taletid.	This is it. It's time for the closing remarks from each party leader.
Vi starter med Anne Sofie Lindenkrone, værsgo.	We'll begin with Sofie Lindenkrone
Tak.	Thank you
Jeg skammer mig over at være dansker…	I'm ashamed to be a Dane…
… når jeg ser, hvad det er for et egoistisk Danmark, Hesselboe og kompagni har vist os de sidste syv år.	… in this selfish nation Hesselboe has created these past seven years
Og Kathrine, det er over til Amir Diwan og det er på 3'eren.	Amir Diwan will be on camera three
Ja, er vi sikre på, at Diwan er med "w"?	Is Diwan spelt with a "w"?
Ja, han er tjekket. Og han er der.	Yes. That's confirmed
En stemme på os, er en stemme på et miljørigtigt…	A vote for us is a vote for a green…
…økonomisk ansvarligt og sundt Danmark	… and financially sound Denmark
Og naturligvis skal Danmark have et forsvar, der er tidssvarende og veluddannet	Denmark, of course, needs a modern, well-trained military…
Så lægger vi om til Hesselboe på 3	Hesselboe on three
Vi tror da også på, at den enkeltes frihed, initiativ og handlemulighed…	Individual freedom and initiative…
… er en vigtig drivkraft for at holde den danske økonomi kørende	… is what keeps the Danish economy going
3 er på der.	Roll three now
Tak, Svend Åge Saltum	Thank you, Svend Age Saltum
Den næstsidste i rækken er Birgitte Nyborg fra de moderate.	Second to last is Birgitte Nyborg.
Ordet er dit.	The floor is yours
Tak.	Thank you
En moderne verden er mangfoldig…	A modern world is manifold…
…og det må vores demokrati også være	…and our democracy should be, as well
En stemme på de moderate i morgen er en stemme på et nyt Danmark. Tak.	A vote for the Moderates tomorrow is a vote for a new Denmark. Thank you

Appendix K

Transcripts of Audio clips used in Experiments 4a and 4b

English	Danish
Passage title: Making and flying a kite	At lave og flyve en drage
A newspaper is better than a magazine.	En avis er bedre end et blad
A seashore is a better place than the street.	En kyst er et bedre sted end gaden.
At first it is better to run than to walk.	Til at begynde med er det bedre at løbe end at gå.
You may have to try several times.	Du kan være nødt til at prøve adskillige gange.
It takes some skill but it's easy to learn.	Det kræver lidt øvelse, men det er let at lære.
Even young children can enjoy it.	Det er sjovt selv for små børn.
Once successful, complications are minimal.	Når først man har lært det, opstår der sjældent problemer.
Birds seldom get too close.	Fugle kommer sjældent for tæt på.
Rain, however, soaks in very fast.	Regnen trænger dog meget hurtigt igennem.
Too many people doing the same thing can also cause problems.	Hvis der er for mange, der laver det samme, kan det give problemer.
One needs lots of room.	Man skal bruge masser af plads.
If there are no complications, it can be very peaceful.	Hvis ikke der opstår problemer, kan det være meget fredfyldt.
A rock will serve as an anchor.	En sten kan fungere som anker.
If things break loose from it, however, you will not get a second chance.	Hvis noget river sig løs fra den, får du dog ikke en chance mere.

English	Danish
Passage title: A trip to space	En rumrejse
Joe looked outside from cramped quarters.	Joe kiggede ud fra den trange plads.
Numerous unknown objects moved swiftly by in vague blackness around his field.	Adskillige ukendte objekter bevægede sig hurtigt forbi i det sorte intet der omgav ham.
Two fearless companions worked along manipulating buttons while reading complex patterns.	To frygtløse makkere samarbejdede om at trykke på knapper mens de læste komplekse mønstre.
Flat familiar homeland now resembled a tiny rubber ball.	Det flade velkendte fædreland lignede nu en lille gummibold.
Everyone here and at home knew that only lifeless things would be found among huge cold mountains surrounding deep barren valleys.	Alle her og derhjemme vidste, at der kun ville blive fundet livløse ting blandt enorme kolde bjerge der omgav dybe golde dale.
But all important papers anxiously awaited their arrival for no man had ever made such big news.	Men alle betydningsfulde aviser ventede ængsteligt på deres ankomst, for ingen før dem havde skabt så store overskrifter.

English	Danish
Passage title: Washing clothes	At vaske tøj
The procedure is actually quite simple.	Proceduren er egentlig ret simpel.
First you arrange things into different groups depending on their makeup.	Først organiserer du tingene i forskellige grupper afhængigt af deres udseende.
Of course, one pile may be sufficient depending on how much there is.	Én bunke kan naturligvis være tilstrækkeligt, afhængigt af hvor meget der er.
If you have to go somewhere else due to lack of facilities that is the next step.	Hvis du bliver nødt til at tage et andet sted hen på grund af manglede faciliteter, er det det næste skridt.
It is important not to overdo it.	Det er vigtigt ikke at overgøre det.
It is better to do too few things at once than too many.	Det er bedre at ordne få ting på én gang fremfor for mange.
This may seem unimportant, but complications from doing too many can easily arise.	Det kan virke trivielt, men der kan nemt opstå problemer, hvis man ordner for mange.
A mistake can be expensive.	En fejl kan være dyr.
The manipulation of the appropriate mechanisms should be obvious.	Mekanismerne og hvordan de påvirkes burde være indlysende.
At first the whole procedure may seem complicated, but soon it will become just another facet of life.	Til at begynde med kan hele proceduren virke kompliceret, men det bliver hurtigt bare endnu en del afl livet.
It is hard to foresee any end to the need for this task.	Det er svært at forestille sig, at behovet for at udføre denne opgave forsvinder.

Appendix L

Supplemental material for Part 1

Gender

Because classifying data by men and women broke random assignment, we checked to make sure there were similar numbers of men and women in each of the four cells.

Table 1.

Experiment 1 N per cell

	Video	No Video
Men	32	40
Women	57	68

Table 2.

Experiment 2 N per cell

	Video	No Video
Men	75	98
Women	203	200

Demand

To address the role of demand, we reviewed subjects' responses to the question asking them what they thought the study was about in Experiment 1. Only 5 of the 198 subjects mentioned the video. For example one subject responded "To see if people could perform an emergency action with only a short video to help guide them". Most of the comments instead mentioned people's confidence in stressful or emergency situations. Excluding subjects who mentioned the video didn't change the pattern of results, if anything

excluding them strengthened the effect (M_{diff} between the video and no video conditions increases from 8.4 to 9.66; p value drops from .04 to .02). These findings suggest that demand was not driving the effects described here.

Within subjects repeated measures results Experiment 2

As an exploratory measure in Experiment 2, we showed the subjects who didn't watch the video prior to making their initial confidence ratings (the no video condition) the video after they had completed the main part of the experiment and asked them to repeat the confidence ratings. We conducted a matched pairs test on these pre- and post-video ratings and found that people were more confident in their ability to land the plane both without dying and as well as a pilot could after watching the video. In other words subjects' confidence ratings to the lower standard significantly increased after watching the video $t(1,300) = 3.55$, $M_{diff} = 4.18$, 95% CI [1.86, 6.50], $p < .001$. Likewise, subjects confidence ratings to the higher standard significantly increased after watching the video $t(1,300)= 2.53$, $M_{diff} = 2.96$, 95% CI [0.65, 5.27], $p = .01$.

Appendix M

Experiment 4 materials norming

The purpose of this experiment was to ensure that people found it easier to comprehend a series of ambiguous passages when they were accompanied by a short title describing the topic. We were also interested in how many Mechanical Turk workers were familiar with these ambiguous passages.

Method

Subjects. We aimed to collect data from 50 MTurk workers. But because MTurk and Qualtrics interact in such a way that it is possible to overcollect a target sample size, we collected data from 52 subjects. We analysed the data from 44 subjects after exclusions (M_{age} = 41.02, SD_{age} = 15.22; 32% identified as men, 68% identified as women; 95% English first language).

Design. We used a between-subjects design with two conditions (title: title, no title).

Procedure. Subjects were randomly assigned to read four difficult-to-comprehend passages in English, either with subtitles or without subtitles. The passages were about driving a car, flying a kite, a trip to space, and washing clothes (Bransford & Johnson, 1972; Wiley & Rayner, 2000). After each passage, subjects answered the following questions. "How difficult was it to comprehend the passage?" (1= very easy, 7 = very difficult). "Have you ever come across this passage before?" (yes, no). "How confident are you that you have seen this passage before?" (1= not at all confident, 7 = very confident)

Results

First, we excluded the data from 8 subjects for providing non-sensical responses. Then to determine how familiar people were with the passages, we measured the proportion of subjects who were confident they had come across each passage before. We pre-registered that if the proportion of subjects who rated confidence >4 was higher than 10%, we would

start the norming process again with different passages. But there were zero exclusions on the basis of the criteria.

Everyone who read the untitled passage about driving a car correctly guessed what it was about. Moreover, those who read the titled version of the passage rated their comprehension as 3.08 on 5-point scale (where lower scores indicate easier comprehension) compared to their untitled counterparts who rated it as 3.32. Because of this trivial difference was markedly less than the comparable difference in the other 3 passages (untitledkite: 4.37; titled kite: 3.56; untitled space: 4.74, titledspace: 3.72; untitledclothes: 4.21, titledclothes: 3.16). Therefore, we dropped this passage from the set of materials and proceeded with a set of three passages.

Compliance questions

Table 1

Compliance questions used in experiments 1a, 1b, 2, 3a, 4a, and 4b.

Compliance question	Experiments used
On the previous page we told you whether the upcoming video would have subtitles. Will the upcoming video have subtitles? (yes, no)	1a, 1b
Could you hear the audio in the video? (yes, no)	1a, 1b
Did you watch the entire video, from start to finish? (yes, no)	1a, 1b, 2, 3a
Did you watch the video more than once? (yes, no)	1a, 1b, 2, 3a
Did you complete this session in a single sitting, without stopping? (yes, no)	1a, 1b, 2, 3a, 4a, 4b
Did you complete the experiment in an environment that is free of noise and distraction? (yes, no)	1a, 1b, 2, 3a, 4a, 4b