The Use of Stereotypical Gender Information in Constructing a Mental Model: Evidence from English and Spanish (1996)

by Manuel Carreiras, Alan Garnham, Jane Oakhill, and Kate Cain

Replication of Experiment 1

by Matthew Libby
LINGUIST 245
Stanford University
Spring 2017

I. INTRODUCTION

What gender do you picture when you think of the word “butcher”? What about “cheerleader”? Or “art historian”? Carreiras et al.’s 1996 paper, “The Use of Stereotypical Gender Information in Constructing a Mental Model,” broadly looks at how gender stereotypes influence comprehension in reading tasks. Namely, it focuses on the way that reading comprehension is “incremental,” and relies on readers integrating information they have previously learned about a given world with new information in the text – and how such integration can be influenced or challenged by pre-existing stereotypes (640). Experiment 1, replicated in this paper, focuses specifically on the “anaphor-antecedent mapping” of pronouns to professions mentioned in prior sentences (643). These professions have been categorized as either male-stereotyped, female-stereotyped, or neutral/non-biased.

What is the role of such latent stereotyped information in the processing of pronouns? If there is a mismatch between the expected pronoun based on stereotype and the pronoun that is actually provided, how does that impact comprehension? Carreiras et al. use reading time as a measure of processing and comprehension. The authors’ prediction is that when we read sentences containing certain professions, we incorporate information about gender stereotypes into our mental representation of the text. Therefore, a mismatch between the stereotypical gender of the profession and the pronoun provided will result in a longer reading time than when there is a match, because mismatches are harder to process cognitively.

II. METHODS

Participants

44 subjects were recruited through Amazon Mechanical Turk, and were paid for their participation. All were within the United States. 5 subjects' data were excluded due to excessive failure of the attention check.

Materials

The selection of professions used in this study was taken from the list of stereotype-weighted professions in the Appendix of the Carreiras et al. paper. This is the way the authors of the original paper described their process of gathering these professions and assigning weights:

> "In order to select professions that were biased either to masculine or to feminine stereotypes or were neutral, 30 subjects, who did not take part in the main reading experiment, were asked to rate 120 professions for 'gender stereotypicality': the likelihood that each would be done by either a man or a woman. It was stressed to subjects that their responses should reflect the way the world is, not how they might feel it should be. … [W]e used the 20 best exemplars from each of the three categories (male, female, and neutral) chosen from the rating study." (463-464).

From these 60 professions, I chose half, 30, to use in my replication (10 from each of the male, female, and neutral categories). This selection was mostly random, but I did pick a few professions because I was interested in if and how the gender stereotype of the profession has changed over time. I only changed two professions from the original list, in both cases due to the fact that the original study was performed by and with speakers of British English: 1) "footballer" (a very highly male-stereotyped profession) was changed the equally male-stereotyped athletic profession in the United States, "football player;" 2) "paediatrician" was changed to the more common American English spelling, "pediatrician." The list of professions used in the original paper, with the professions used in this replication highlighted, is provided as Appendix 1.

As in the original paper, from these professions, two-line texts were created. The texts used in the 1996 study were not included in their paper, so 30 original texts were written, one for each profession being used in the replication. Many of these texts were adapted from a similar study conducted by Kennison & Trofe (2003). The newly-created texts followed the same structure as the original study: The first sentence introduced a character only by profession (either male-stereotyped, female-stereotyped, or neutral). The second sentence qualified the first sentence, and crucially referred to the aforementioned character anaphorically (in the case of the stereotyped professions, either matching or mismatching the stereotype). These sentences were followed by a simple question which acted as both an attention check and a way to distract subjects from what was being tested. A full list of two-line texts used in the replication, with accompanying questions, is provided as Appendix 2.

Procedure

(Click this link to view the experiment.)

The experiment was conducted over Amazon Mechanical Turk. Each subject read 30 experimental texts. 10 texts contained professions from the male-stereotyped category, 10 contained professions from the female-stereotyped category, and 10 contained professions from the neutral category. For each category, half of the texts were randomly selected to have a masculine pronoun in their second sentence, and the other half had a feminine pronoun in their second sentence.

In each trial, the two sentences and question were presented successively in the center of the screen. Before each text, subjects hit the SPACE bar to display an initialization prompt:

"Press the SPACE bar to begin." Hitting the SPACE bar again displayed the first sentence, while successive hits displayed the second sentence and then the question. Subjects had to answer "Yes" or "No" to the question to continue. As in the original experiment, "The instructions stressed that the sentences were to be read at normal reading speed, but that the questions were to be answered as quickly and as accurately as possible" (644). Before the presentation of the experimental materials, three practice trials were presented to introduce subjects to the self-paced reading procedure.

Discussion of Differences to the Original Study

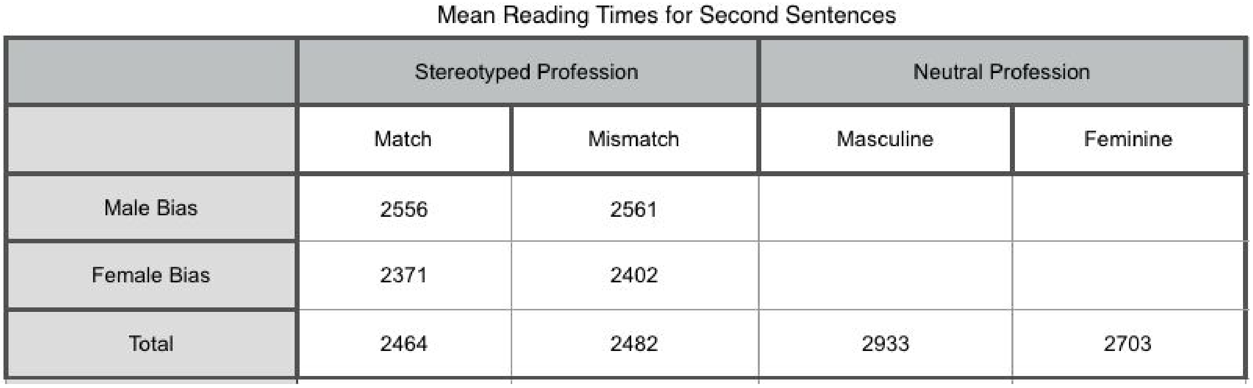
There are a number of differences between this replication and the original study. My replication was run on Amazon Mechanical Turk, whereas the original study was conducted in person. While the original study only used 24 participants (all living in the United Kingdom), the online nature of my replication allowed me to almost double the amount of participants to 44 (after exclusions, 39; all living in the United States). For the sake of keeping the online HIT a reasonable length for MTurk users, I decided to halve the amount of practice and experimental trials. While the original study had 60 experimental and 6 practice texts, I had 30 and 3, respectively. Other changes already mentioned include my creation of original two-line texts and questions due to lack of information about them in the original study, as well as my changing certain British words and spelling to their appropriate American counterparts.

The only major experimental design change I made was that I decided to exclude filler trials from the replication. The original study contained 24 filler items, which "were included to reduce the proportion of texts about people whose occupations did not match the gender stereotype, so that subjects would not realize that these texts were under investigation" (644). However, I decided not to include them for two reasons: 1) I was minding the length of the online HIT, and 2) I figured the neutral texts would play the same role in “reducing the proportion of texts about people whose occupations did not match the gender stereotype.” For these reasons, filler items were not included.

III. RESULTS

After exclusions, the vast amount of attention check questions were answered correctly. In total, 1094 questions were answered correctly, while only 70 questions – across 39 participants – were answered incorrectly.

The mean reading times for the second sentences of the replication trials are shown in Figure 1.



Mean Reading Times for Second Sentences

	Stereotyped Profession		Neutral Profession	
	Match	Mismatch	Masculine	Feminine
Male Bias	2556	2561		
Female Bias	2371	2402		
Total	2464	2482	2933	2703

Figure 1. Mean Reading Times for Second Sentences

Note that mean reading times reflect essentially no difference between the match and mismatch conditions of the stereotyped professions, but a noticeable difference between the masculine and feminine pronoun conditions of the neutral professions. Also note that the male bias results in a longer reading time for both the match and mismatch conditions. Finally, and most importantly, note that the neutral profession trials resulted in reading times almost half a second longer than the stereotyped professions.

Figure 2 shows this table in graph form, with error bars.

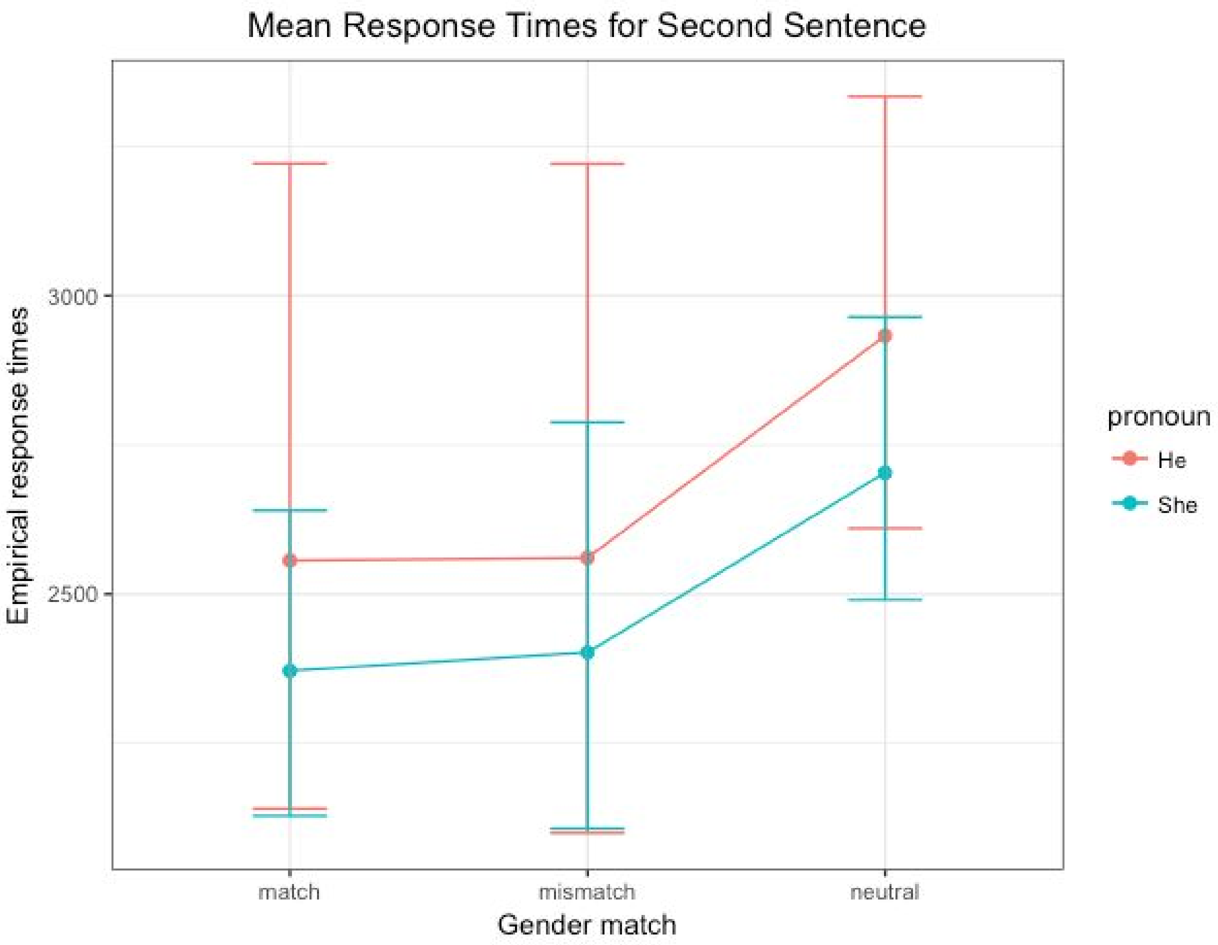


Figure 2. Mean response times for second sentence.

The uncertainty implied by the substantial error bars in Figure 2 can be further dissected by looking at response time by profession, shown in Figure 3. For most stereotyped professions, there appears to be no significant difference between the match and mismatch conditions, nor between the masculine and feminine conditions of the neutral professions. In the cases that there does seem to be a noticeable difference (e.g. “art historian,” “cleaner,” “football player”), that difference is accompanied by sizable error.

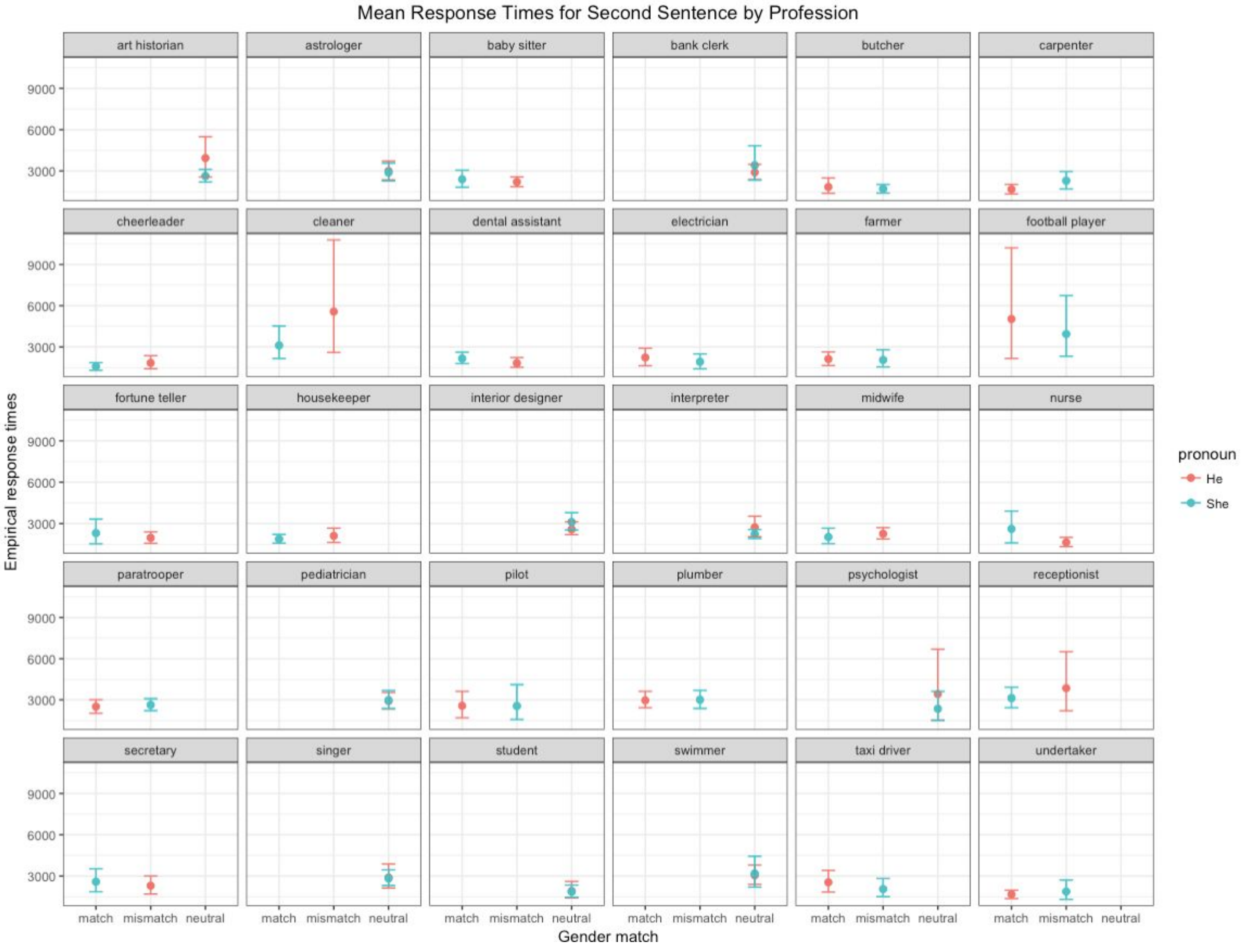


Figure 3. Mean Response Times for Second Sentence by Profession.

These intuitions are borne out in analysis. While the original paper reported analyses generally derived from 2x2 ANOVAs, I translated these ANOVAs to mixed effect linear regression models (each with subject and/or profession as random variables). In one model, which regresses log-transformed reading time onto the interaction of gender match and the stereotyped/expected pronoun, I found no significance in the main effects of gender match (p = .674) and stereotyped pronoun (p = .937), nor their interaction (p = .935). In another model, which factored in by-word sentence length as well, I found no significance in its main effect (p = .454) nor its interaction with the above fixed variables (p = .777). (The previously mentioned variables and their interaction retained their lack of significance in this second model.) Finally, a model which solely looked at the reading times of the neutral profession trials as an effect of the pronoun used found the difference between reading times in the masculine and feminine pronoun conditions to be not significant (*p* = .337).

The non-significance of the main effect of the stereotyped pronoun, the interaction of gender matching and the stereotyped pronoun, and the effect of pronoun gender on neutral reading times were all found in the original study. The original study did find the main effect of gender significant. That is, they found that the mismatch condition resulted in significantly longer reading times than the match condition, confirming the authors' hypothesis and prompting them to write "a mismatch with a stereotype slowed people down considerably" (646). My replication did not find this effect significant, however, confirming what can be intuited from Figure 1. Finally, the original study controlled for sentence length by character, and did find a significant difference between the mismatch and neutral conditions in doing so. My replication, on the other hand, found no effect in any condition when controlling for sentence length by word.

One significant effect was found, however, when running a model not adapted from the ANOVAs run for the original paper. In a simpler model which simply regresses log-transformed reading time onto the match or mismatch of the pronoun, significance was found in the reading times of neutral trials being longer than the reading times of either the match or mismatch trials (*p* = .031). This is surprising, as this result was not even observed in the original paper (the authors found that, on average, neutral reading times were longer than match reading times but shorter than mismatch reading times), and it was the only significant effect found in the replication.

IV. DISCUSSION

The motivating question behind both the original study and this replication was whether the match or mismatch of an gendered anaphor to a gender-stereotyped antecedent would impact reading comprehension, using reading time as a measure of processing. While the original paper found a significant increase in reading time in the mismatch condition, this replication found no such effect. In fact, the replication found a significant increase in reading time for the *neutral* professions. The replication did agree with the finding from the original study that there is no significant difference between reading time of masculine and feminine pronouns when referring to neutral professions.

As for the main difference between these results and the results of the original study – that being the result regarding the significance of longer neutral reading times – it suggests that it might be more difficult for subjects to map an anaphor to an antecedent when there *isn't* any gender-stereotype associated with the antecedent. However, there are two aspects of my experimental design that I believe could've contributed to the difference. Firstly, the absence of filler items in this replication – it is possible that neutral items didn't end up playing the role of fillers in helping to distract subjects from the pattern of the experiment. If subjects did subconsciously begin to realize that many professions they were being asked about had a strong gender stereotype, then they might have been surprised by the neutral professions. This surprise could have contributed to a longer reading time, but only because these trials were in the context of the stereotyped professions.

Secondly, it is also possible that some of the neutral professions are not neutral in their stereotyping twenty years after the original study. That is, certain professions that were stereotyped as neutral in 1996 Britain might not be stereotyped as neutral in 2017 America – the categories as they were delineated by the original authors may have shifted, skewing the overall results. However, it is difficult to make any definitive claims because a) I did not redo the “gender stereotypicality” task that was done for the original study, and simply used their professions and weights, and b) the original paper does not provide any by-profession reading time data, meaning I cannot make any assessment as to how mean reading times have changed for individual professions over the past twenty years.

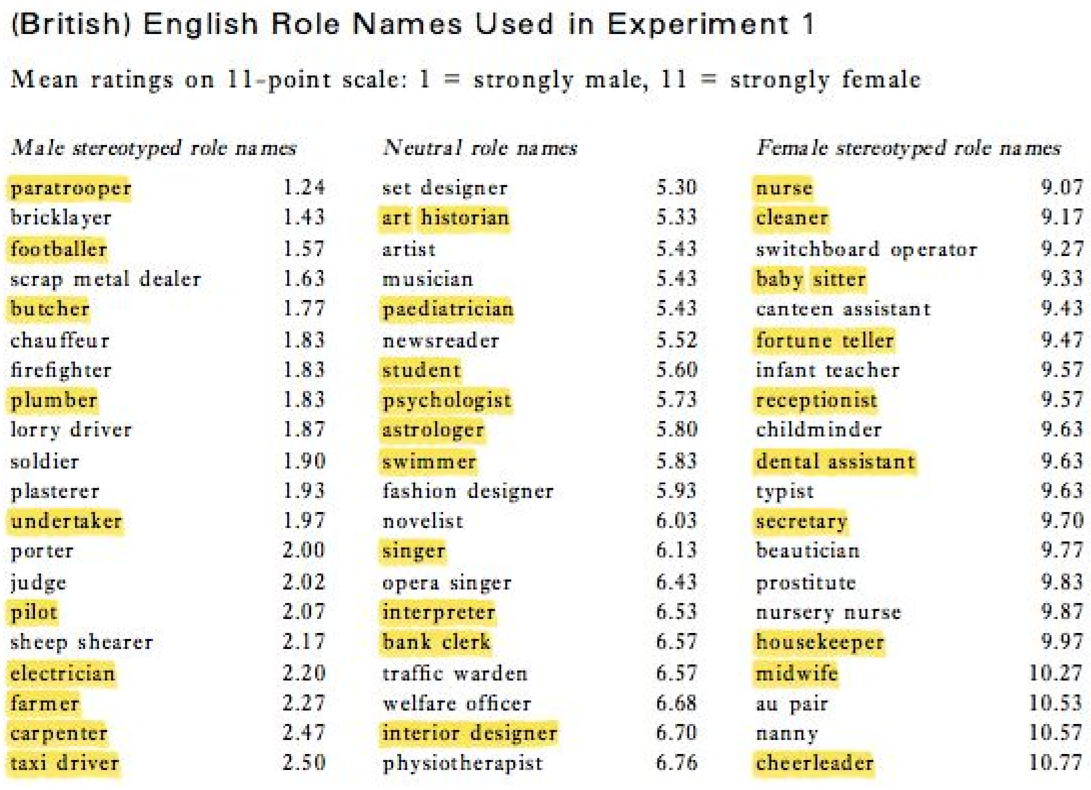
Overall, the results of this replication disagree with the major finding of the original paper by Carreiras et al., namely that a mismatch between the gender of an anaphor and the stereotyped gender of its antecedent will impact comprehension in reading tasks. Instead, the replication found that anaphors pointing to neutral antecedents are harder to process than those pointing to gender-stereotyped antecedents, and that there is no significant difference in comprehension between match and mismatch conditions. While there are plausible explanations for these differences in the design of the replication, making any definitive claim would require further experimentation.

V. REFERENCES

Carreiras, Manuel, Alan Garnham, Jane Oakhill, and Kate Cain. "The Use of Stereotypical Gender Information in Constructing a Mental Model: Evidence from English and Spanish." *The Quarterly Journal of Experimental Psychology* (1996): 639-63. Web.

Kennison, Sheila M., and Jessie L. Trofe. "Comprehending Pronouns: A Role for Word-Specific Gender Stereotype Information." *Journal of Psycholinguistic Research* 32.3 (2003): 355-78. Web.

APPENDIX 1



(British) English Role Names Used in Experiment 1

Mean ratings on 11-point scale: 1 = strongly male, 11 = strongly female

Male stereotyped role names		*Neutral role names*		*Female stereotyped role names*	
paratrooper	1.24	set designer	5.30	nurse	9.07
bricklayer	1.43	art historian	5.33	cleaner	9.17
footballer	1.57	artist	5.43	switchboard operator	9.27
scrap metal dealer	1.63	musician	5.43	baby sitter	9.33
butcher	1.77	paediatrician	5.43	canteen assistant	9.43
chauffeur	1.83	newsreader	5.52	fortune teller	9.47
firefighter	1.83	student	5.60	infant teacher	9.57
plumber	1.83	psychologist	5.73	receptionist	9.57
lorry driver	1.87	astrologer	5.80	childminder	9.63
soldier	1.90	swimmer	5.83	dental assistant	9.63
plasterer	1.93	fashion designer	5.93	typist	9.63
undertaker	1.97	novelist	6.03	secretary	9.70
porter	2.00	singer	6.13	beautician	9.77
judge	2.02	opera singer	6.43	prostitute	9.83
pilot	2.07	interpreter	6.53	nursery nurse	9.87
sheep shearer	2.17	bank clerk	6.57	housekeeper	9.97
electrician	2.20	traffic warden	6.57	midwife	10.27
farmer	2.27	welfare officer	6.68	au pair	10.53
carpenter	2.47	interior designer	6.70	nanny	10.57
taxi driver	2.50	physiotherapist	6.76	cheerleader	10.77

The original list of professions and their gender-stereotyped weights from the original Carreiras et al. paper (1996). Highlighted are the professions used in this replication.

APPENDIX 2

A list of the 30 experimental items – two-sentence texts and a question (attention check):

The art historian gets many invitations to speak at college campuses.
[He/she] only accepts a few in the New England area.
Does the art historian ever speak outside of the New England area?

The astrologer got a sweepstakes entry in the mail.
[He/she] decided to fill it out and mail it in.
Did the astrologer receive a package in the mail?

The babysitter found out about the practical joke.
[He/she] remained calm for several minutes, but then started to yell at everyone.
Did the babysitter react poorly to the practical joke?

The bank clerk had worked in that town for over thirty years.
[He/she] knew just about everyone who lived there.
Did the bank clerk know just about everyone in the town?

The butcher seemed a little distracted.
[He/she] said that there were several family members who were in the hospital.
Was the butcher happy?

The carpenter walked to the grocery store.
[He/she] didn’t make eye contact with anyone passing by.
Did the carpenter make eye contact with passersby?

The cheerleader arrived early for the game
[He/she] was amazed to see how many people were there.
Did the cheerleader arrive late for the game?

The cleaner read the instructions on the bottle very carefully.
[He/she] then began to mix the cleaning solution.
Was the cleaner mixing a cleaning solution?

The dental assistant made several recommendations.
[He/she] stressed the importance of avoiding sugary snacks.
Did the dental assistant encourage avoiding sugary snacks?

The electrician examined the light fixing.
[He/she] needed a special attachment to fix it.
Was the electrician mending a stereo?

The farmer was worried about the next harvest.
[He/she] was concerned there wouldn’t be enough rain.
Was the farmer concerned there wouldn’t be enough rain?

The football player spent hours in the whirlpool each day.
[He/she] had injured an ankle, and the whirlpool eased the pain.
Had the football player injured an ankle?

The fortune teller traveled with the carnival.
[He/she] had been born into the carnival life and never wanted to leave it.
Had the fortune teller joined the circus as a teenager?

The housekeeper always showed up late for work.
[He/she] was threatened with termination if this behavior continued.
Did the housekeeper always show up for work on time?

The interior designer inspected the ballroom before the gala.
[He/she] found everything in perfect order.
Was the interior designer inspecting the garden?

The interpreter checked the documents carefully for typos.
[He/she] signed them and put them in the outgoing mail.
Did the interpreter put the documents in the trash?

The midwife read the newspaper everyday.
[He/she] tried to keep informed about national and world news.
Did the midwife read the newspaper only once a week?

The nurse examined the medication carefully.
[He/she] then wrote something on the chart and left.
Was the nurse examining medication?

The paratrooper was very moody and hard to get along with.
[He/she] seemed to be liked by no one in the squad.
Did anyone in the squad like the paratrooper?

The pediatrician gave out free gift bags to every appointment.
[He/she] included in each gift bag the child's favorite candy.
Did the pediatrician give out free gift bags to only some appointments?

The pilot announced the time and the weather.
[He/she] indicated that the plane would be landing a little ahead of schedule.
Was the plane going to arrive late?

The plumber stopped by the diner on the way home from work.
[He/she] ordered the special: catfish sandwich and fries.
Was the special a catfish sandwich?

The psychologist liked to visit the zoo.
[He/she] found watching the animals to be a perfect way to relax after work.
Did the psychologist like to visit the zoo after work?

The receptionist received an invitation to a charity benefit.
[He/she] decided not to go because there was a lot piling up at work.
Did the receptionist receive an invitation to a charity benefit?

The secretary distributed an urgent memo.
[He/she] made is clear that work would continue as normal.
Was the memo the secretary distributed urgent?

The singer spent hours every week watching television.
[He/she] would watch even more if it were possible.
Did the singer like watching television?

The student seemed upset and pale.
[He/she] then asked for a glass of water.
Did the student ask for a glass of water?

The swimmer attended a press conference before the latest competition.
[He/she] said that there was every reason to expect a gold medal.
Was the swimmer at a press conference?

The taxi driver took the twenty dollar bill.
[He/she] immediately thought it might be counterfeit.
Was the taxi driver handed a five dollar bill?

The undertaker went outside to smoke.
[He/she] finished smoking two cigarettes before returning inside.
Did the undertaker smoke three cigarettes?