Working memory updating modulates adaptation to speaker-specific use of uncertainty expressions

Sebastian Schuster, **Alexandra Mayn** and **Vera Demberg**
{seschust, amayn, vera}@lst.uni-saarland.de
Department of Language Science and Technology, Saarland University
66123 Saarbrücken, Germany

Abstract

Listeners rapidly learn speaker-specific expectations and interpretations of words and phrases such as uncertainty expressions when they observe a speaker's use of these expressions. However, previous studies have exclusively examined this behavior in *populations* of listeners and it remains unclear to what extent there are systematic individual differences in listeners' adaptation behavior and, if such differences exist, whether they are linked to more general cognitive abilities. In this work, we first re-analyze the data by Schuster and Degen (2019) and show that listeners vary in the extent to which they adapt to different speakers. In a series of exploratory and confirmatory studies, we then show that the extent to which listeners update their expectations of different speakers is correlated with participant's score on the Keep Track Task (Yntema, 1963), which suggests that working memory control modulates listeners' semantic-pragmatic adaptation abilities.

Keywords: adaptation; individual differences; working memory; pragmatics; psycholinguistics; uncertainty expressions; language comprehension

Introduction

A hallmark of language processing is the ability of listeners to *adapt* to individual speaker's language use at all linguistic levels (Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2007; Bradlow & Bent, 2008; Kleinschmidt & Jaeger, 2015; Fine & Jaeger, 2016; Roettger & Franke, 2019; Xiang, Kramer, & Kennedy, 2020, *i.a.*). This includes adaptation at the semantic-pragmatic level: When listeners are exposed to two different speakers who vary in their use of quantifiers such as *some* and *many* or uncertainty expressions such as *might* and *probably*, listeners rapidly develop speaker-specific expectations about the use of these expressions that allow them to make more precise interpretations (Yildirim, Degen, Tanenhaus, & Jaeger, 2016; Schuster & Degen, 2019, henceforth S&D). For example, S&D exposed participants to two speakers who varied in their use of the uncertainty expressions *might* and *probably* to communicate probabilities of an uncertain event. One of the speakers, the "cautious" speaker always produced *might* for event probabilities of 60%, whereas the other speaker, the "confident" speaker always produced *probably* in that same situation. On filler trials the speaker used the other uncertainty expression for probabilities which received very high prior ratings or the *bare* assertion "You will get a blue one" (see Table 1 for a summary of the exposure trials). They found that when participants were probed for their expectations of the use of uncertainty expressions after exposure to these two speakers, listeners provided different responses for the two speakers and the responses were closely aligned with the behavior that participants saw during the exposure phase, suggesting that listeners learned speaker-specific expectations.

Up until this point, research into such semantic-pragmatic adaptation behavior has focused on populations of listeners, and S&D and Yildirim et al. (2016) only reported results at a population level. At the same time, however, there is increasing evidence that there exist systematic individual differences in pragmatic behavior for a range of phenomena, such as deriving scalar and ad-hoc implicatures (Franke & Degen, 2016; Yang, Minai, & Fiorentino, 2018; Mayn & Demberg, 2022), comprehension of indirect requests and metaphors (Fairchild & Papafragou, 2021; Trott & Bergen, 2019), and drawing coherence-relation inferences (Scholman, Demberg, & Sanders, 2020), and that these individual differences are in part explained by individual differences in general cognitive capacities such as theory-of-mind (ToM) or the amount of linguistic experience.

In this work, we build upon these recent findings, and investigate to what extent similar systematic individual differences exist in the capacity of semantic-pragmatic adaptation, specifically in the domain of uncertainty expressions. We first perform a reanalysis of the results by S&D using a novel measure for individual adaptation behavior and find that there are considerable differences across participants in the extent to which participants adapted to the two speakers in their experiment. We then conduct an exploratory experiment to evaluate to what extent four different individual difference measures predict how much individual participants adapt to the two speakers (Exp. 1). Specifically, we consider measures which aim to assess memory updating, theory of mind, and cognitive reflection abilities as well as a measure assessing language experience. We find that only performance on the keep track task (Yntema, 1963), which aims to measure memory updating capacities, predicts the magnitude of adaptation. We replicate this exploratory finding in a pre-registered confirmatory experiment (Exp. 2).

Data and models are available at `https://github.com/sebschu/adaptation-id`.

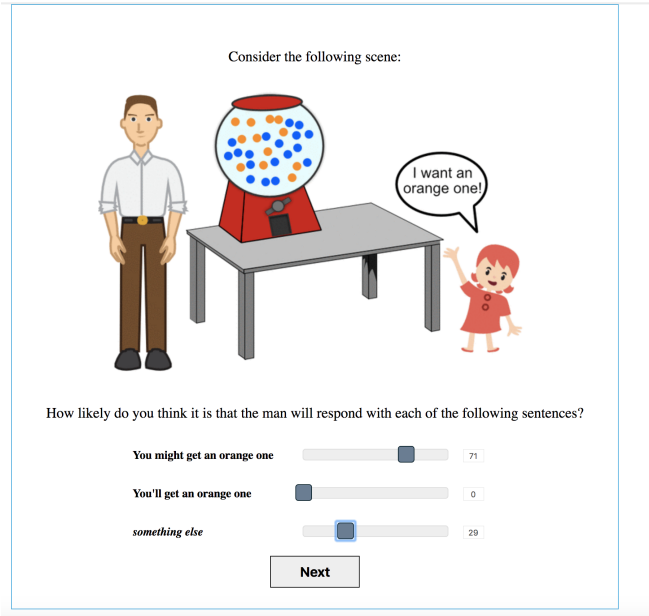

Figure 1: Example trial from prior elicitation experiment.

Measures

Before reporting our experiments, we briefly describe the measures we used to estimate adaptation behavior as well as the general cognitive measures that we collected.

Adaptation experiment

We use the gumball paradigm from S&D. In this paradigm, participants see a drawing of a child standing next to a gumball machine with orange and blue gumballs in it, and an adult standing on the other side of it (see Fig. 1). Participants are told that the gumball machine is too high up for the child to see. The proportion of gumballs of each color in the gumball machine varies by trial. Schuster and Degen (2020) collected participants' prior beliefs about a generic speaker's use of uncertainty expressions by asking participants to distribute 100 points between utterances containing the two uncertainty expressions *might* (*You might get a blue one*) and *probably* (*You'll probably get a blue one*) and a blanket *something else* option, depending on the ratio of orange and blue gumballs in the machine. They observed that there were probability ranges (at around 60%) for which on average participants assigned similar numbers of points to *might* and *probably*, indicating that listeners have uncertainty about what expression a speaker might use. These priors are then averaged to serve as the population prior in a Bayesian adaptation model.

In S&D, the paradigm is extended to test adaptation to multiple speakers: each participant is exposed to two speakers in a blocked design, one of whom is "cautious" and the other is "confident." The authors find that within participants there is also significant difference between post-exposure ratings for the two speakers, with higher ratings being assigned to *probably* for the "confident" speaker, although the effect size is smaller compared to the one-speaker experiment.

	might		*probably*		*bare*	
	n	p	n	p	n	p
cautious	**10**	**60%**	5	90%	5	100%
confident	5	25%	**10**	**60%**	5	100%

Table 1: Number of exposure trials (n) per utterance (MIGHT, PROBABLY, BARE) and associated proportion of target gumballs (p) in the cautious vs. confident speaker block. Critical trials bolded.

For measuring adaptation behavior, we use the same experiment as S&D and collect data in a test-exposure-test experiment, which consists of prior belief elicitation, exposure to two speakers in a blocked design, and collecting post-exposure uncertainty ratings for each speaker.[1]

Our prior elicitation design is identical to Exp. 1 of Schuster and Degen (2020). While they observed variability in the prior ratings, they only used that data to obtain an averaged population-level prior for a generic speaker. In the current study, we collect per-participant prior ratings since we believe that those are likely to be relevant to adaptation: where one's beliefs start out influences the direction and amount of updating one has to do. Also, since we are interested in the effect of individual differences on adaptation, possible effects of prior should be accounted for by adding it into the statistical model as a covariate.

The exposure and posterior elicitation follow Experiment 1 of S&D: the cartoon man from prior elicitation is replaced with a video of a man or a woman, one of whom is a "confident" and the other a "cautious" speaker (counterbalanced assignment). Participants are then exposed to 20 utterances by each speaker (see Table 1) in a blocked design, with counterbalanced block order. Finally, they are asked to rate the likelihood of each of the speakers uttering *might* and *probably* or a *something else* option by distributing 100 points between those responses for different proportions of gumballs.

In addition, to ensure that participants are paying attention to the visual scenes, we include attention checks: After 14 of the 40 exposure trials, participants are shown two machines with different proportions of blue gumballs and asked to click on the one they saw on the previous trial.

[1] Schuster and Degen (2020) collected prior ratings in a separate experiment from the main adaptation experiment. As we discuss below, we wanted to estimate to what extent listeners' prior beliefs affect post-exposure adaptation behavior and therefore, we collected prior beliefs as part of the main adaptation experiment. In theory, it could be that asking participants to provide ratings before the exposure phase increases their awareness about what the experiment is about and consequently affects their post-exposure behavior. To rule out such an effect, we also ran an exact replication of the S&D adaptation experiment and found that at a population level, post-exposure ratings were almost identical independent of whether the experiment included a prior elicitation task or not ($r = 0.996$).

Measure S&D's main dependent measure was the difference in the area under the curve (AUC) between the ratings for the sentence with *might* and the sentence with *probably*, which captures the finding that participants expect the speaker to use *might* for a larger range of event probabilities in the "cautious" speaker condition than in the "confident" speaker condition, and that participants expect the speaker to use *probably* for a smaller range of probabilities in the "cautious" condition. While this measure works well for comparing participants' behavior at the population level, we found that this measure is too noisy to draw useful conclusions about individual participants' behavior. We therefore derived a measure from the computational model by Schuster and Degen (2020).

The model by Schuster and Degen (2020) is an instance of an *ideal observer* model (Kleinschmidt & Jaeger, 2015). According to this model, listeners have beliefs about how a specific speaker uses uncertainty expressions which depend on beliefs about the speaker's meaning of uncertainty expressions and beliefs about the speaker's preferences. In interaction, when a listener observes how a speaker uses uncertainty expressions, listeners update their beliefs through Bayesian belief updating and refine their expectations about the speaker's use and consequently also their interpretations of the speaker's utterances. Schuster and Degen (2020) estimated priors for this model and then simulated the adaptation process by combining the priors with the observed utterance-event probability pairs to obtain models of listeners' beliefs after being exposed to the "cautious" or "confident" speaker, respectively, and they showed that such models of posterior beliefs closely predict participants' behavior after exposure.

To analyze individual participants' behavior, we use the posterior models for the "cautious" and "confident" speakers. Specifically, we compute the likelihood of each participant's responses in the "cautious" speaker and "confident" speaker condition under both models. This tells us whether a participant's behavior in a specific condition resembles more the expectations of a "cautious" or "confident" speaker. Further, we can compute the log likelihood ratio of a participant's ratings for a specific condition between the two models to estimate how much more the data resembles one of the two speaker types.

To verify that this measure leads to similar results as the AUC measure, we re-ran the regression model by S&D with the log likelihood ratio (LLR) as a dependent measure. Similarly as S&D, we found that only condition is a significant predictor of LLR (Model predicting LLR: $t(139) = 2.46$, $p < 0.05$; Model predicting AUC: $t(139) = 2.91$, $p < 0.01$), thus replicating the results using our model-based measure.

Individual difference measures

Keep Track Task (KTT) Yntema (1963) proposed the KTT as a measure of working memory updating, which is the ability to maintain and modify representations in working memory. It is related to and correlated with working memory capacity but is distinct from it (Ecker, Lewandowsky, Oberauer, & Chee, 2010; Frischkorn, Von Bastian, Souza, & Oberauer, 2022). This task has been widely used in studies investigating executive function and its relationship with reading and reasoning (Friedman et al., 2006; Johann, Könen, & Karbach, 2020; McIlhiney, Gignac, Ecker, Kennedy, & Weinborn, 2022).

To adapt to speaker-specific language use, listeners need to accurately store how a specific speaker used language in the past. In addition, when one is exposed to multiple speakers, one also needs to keep representations of the two speakers separate in memory and update these representations correctly. Therefore, we speculated that participants with better working memory control may be better at keeping track of which uncertainty expression came from which speaker, resulting in a greater adaptation effect.

On each trial, participants were told to keep track of two to four out of six possible categories and at the end recall the last word in each tracked category. The dependent measure was the number of correctly recalled words.

Cognitive Reflection Test (CRT) was proposed by Frederick (2005) to measure reflexivity, or how likely a person is to reflect on their first intuitive response. Performance on the CRT has been shown to correlate with consistency in scalar implicature comprehension (Heyman & Schaeken, 2015) and with pragmatic responding in a reference game (Mayn & Demberg, 2022).

We speculated that participants may need to inhibit the response corresponding to their own prior beliefs about the meaning of words in order to provide a response reflecting other speakers' distinct use of uncertainty expressions.

We used the 10-question version of CRT used in Mayn and Demberg (2022), with 6 critical questions and 4 decoy questions, selected from existing versions of CRT (Primi, Morsanyi, Chiesi, Donati, and Hamilton (2016); Baron, Scott, Fincher, and Metz (2015); Sirota and Juanchich (2018); Thomson and Oppenheimer (2016); Toplak, West, and Stanovich (2014)). The score is the proportion of correctly answered previously unseen critical questions. Participants who reported having seen 3 or more of the 6 critical questions were excluded from analysis.

Author Recognition Test (ART) The ART (Stanovich & West, 1989) measures participants' exposure to print and has been used as a proxy for participants' linguistic experience (Scholman et al., 2020; Johnson & Arnold, 2021). In this task, participants are presented with 130 names in alphabetical order, half of which were real author names and the other half were foils, and responded for each name whether it is an author name or not. Participants are also instructed not to guess. The score is the number of correctly identified real author names minus the number of falsely identified foils.

We speculated that participants with more linguistic experience might be either more open to the idea that words can be used in different ways by different speakers or have stronger prior expectations about the meaning of words and therefore

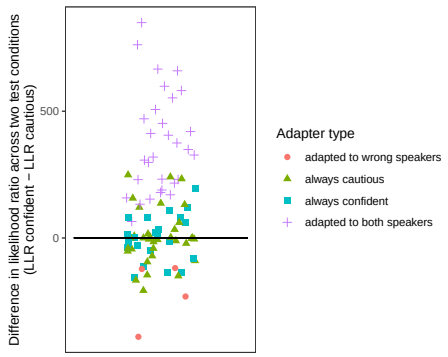

Figure 2: Results of reanalysis of individual participant behavior in the experiment by S&D.

explored whether there would be a larger or a smaller adaptation effect for individuals with higher reading experience.

Reading Mind in the Eyes (RMET) RMET (Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997) is a commonly used measure of Theory of Mind, which can be defined as the tendency to reflect on others' beliefs and desires and to allow for the possibility that they are different from one's own. Participants with higher ToM, or mentalizing, ability have been shown to be more pragmatic in their comprehension of scalar implicatures, metaphors and indirect requests (Fairchild & Papafragou, 2021; Trott & Bergen, 2019). We opted for RMET because it is the most commonly used and easily administered ToM test that does not result in ceiling effects in the neurotypical population, and it is highly correlated with other ToM measures, such as the Short Story Task (Dodell-Feder, Lincoln, Coulson, & Hooker, 2013).

In this task, participants see 36 pictures where only eyes of a person are visible and have to identify the emotion that person is experiencing, choosing from one of 4 options. The score is the number of correctly answered questions.

We speculated that participants who are more likely to reason about others' beliefs may be more open to the idea that a speaker may have different beliefs about the meaning of uncertainty expressions, resulting in a greater adaptation effect.

Individual differences in the data by S&D

As a starting point for our investigation, we conducted a reanalysis of the results of Experiment 1 in S&D[2] using the model-derived measure to determine whether there exist systematic differences in the behavior of individual participants.

Figure 2 shows the difference in LLR across conditions for each participant. As these results show, participants vary in terms of how much their expectations diverged across the two speakers that they saw during the exposure phase, with many participants providing very similar expectations across conditions as indicated by the many data points close to 0. Crucially, however, there is also a group of participants who provided very different speaker expectations across conditions. This suggests that the majority of participants had similar expectations for both speakers and did not learn speaker-specific expectations but a smaller group of participants seemed to adapt to both speakers. We further classified participants into four categories depending on whether their behavior differed across conditions but was the opposite from what was expected (*adapted to wrong speakers*), whether they always provided responses most similar to a "cautious" speaker or always most similar to a "confident" speaker (*always cautious* or *always confident*), or whether they actually aligned their expectations to the behavior of both speakers (*adapted to both speakers*), which further confirmed that almost all participants either primarily adapted to one of the two speaker types or learned speaker-specific expectations for both speakers. Subsequently, we explore to what extent these differences can be predicted from the measures discussed above.

[2]Data available at https://github.com/sebschu/adaptation.

Experiment 1: Exploratory study

The goal of this experiment was to first replicate the results of S&D and to investigate whether individual differences in cognitive traits modulate adaptation.

Participants

130 native English speakers born and residing in the United States were recruited via the crowdsourcing platform Prolific. Participants were paid $11 (∼$12/hr).

Procedure

Participants completed the adaptation task followed by the battery of individual difference tests in the following order: KTT, RMET, CRT and ART. We first recruited 65 participants to complete only the adaptation task and then re-invited participants for the second session with individual differences tasks, which were completed by 37 participants. Due to the relatively low return rate, we asked the second batch of 65 participants to complete the adaptation experiment and the individual difference tasks in one session.

Results

A total of 100 participants completed all tasks. We excluded 2 participants who answered more than 3 of the 14 attention checks wrong. We excluded 3 additional participants for performing below 2.5 SDs below the mean on KTT; 13 additional participants for random guessing on the ART task; 14 due to familiarity with the CRT task; and 1 for performing below 2.5 SDs below the mean on RMET, resulting in 67 participants for whom we had all four individual difference measures.[3] Summary statistics for the individual difference measures are reported in Table 3. There was a moderate positive correlation between KTT and RMET (Pearson's r=0.36, p=0.003, Bonferroni-Holm-corrected) but none of the other correlations between tasks reached significance.

[3]Considering this large number of exclusions, we also performed all analyses on the data of all 98 participants and found that none of the findings reported here hinge on the exclusions.

	Exp. 1: Full model ($n = 67$)			Exp. 1: Reduced ($n = 95$)			Exp. 2 ($n = 91$)		
	β	SE	p	β	SE	p	β	SE	p
Intercept	−6.65	27.29	0.808	−6.71	17.57	0.703	−31.08	16.00	0.052
Condition	−84.54	16.06	**1.40e−07**	−82.99	12.92	**1.32e−10**	−58.40	11.85	**8.27e−07**
Test order	−16.91	21.68	0.435						
Most recent speaker	−53.47	22.26	**0.016**	−49.14	17.59	**0.005**	−50.16	16.02	**0.002**
Prior likelihood ratio	54.70	55.02	0.320						
KTT	−5.90	47.09	0.900	26.18	36.30	0.471	48.90	36.48	0.180
ART	−106.83	52.50	**0.042**						
RMET	80.76	61.61	0.190						
CRT	28.11	38.36	0.464						
Condition:KTT	−78.49	35.31	**0.026**	−94.17	26.68	**4.15e−04**	−77.95	26.98	**0.004**
Condition:ART	2.06	37.13	0.956						
Condition:RMET	−34.58	44.85	0.441						
Condition:CRT	−27.68	28.91	0.338						

Table 2: Model statistics.

Measure	Mean	SD	Obs. range	Poss. range
ART	23.13	11.94	2-59	-65-65
CRT	0.26	0.27	0-1	0-1
RMET	28.21	3.49	19-35	0-36
KTT	28.42	3.65	20-35	0-36
KTT	28.70	3.31	20-35	0-36

Table 3: Statistics of individual difference measures collected in Experiment 1 (top part) and Experiment 2 (bottom part).

We fit a linear mixed-effects model to replicate the adaptation effect to the two different speakers from S&D and investigate the relationship between adaptation and individual differences. Our dependent measure is the log likelihood ratio obtained by computing the likelihood of each participant's responses in both conditions under both the cautious and the confident models by Schuster and Degen (2020) (see above for details). We regress the log likelihood ratio onto condition (1: cautious vs. -1: confident), test order (whether participants provided post-exposure responses for the two speakers in 1: the same order as they saw them during exposure or -1: not), prior log likelihood ratio, which speaker they saw last (most recent speaker type, 1: cautious or -1: confident), and interactions of each of the individual differences (scaled to between -1 and 1 and centered) with condition. The model also included random per-subject intercepts. Note that we are interested in the interaction between the individual differences with condition and not the main effect of the individual differences because adapting to two different speakers means an increase or a decrease of the log likelihood ratio depending on the condition.

The full model is reported in the first column of Table 2. Consistent with S&D, there was a main effect of condition ($\beta = -84.54$ (16.06), $p < 0.001$), suggesting that at the population level, participants adapted to both speakers. There was also a recency effect of which speaker the participants saw before the exposure block ($\beta = -53.47$ (22.26), $p < 0.05$): if they saw the cautious speaker second, participants behaved more similarly to the cautious speaker in both test blocks, and vice versa.

Of the individual differences, we only observed a significant interaction of condition with the keep track task ($\beta = -78.49$ (35.31), $p < 0.05$), which measures working memory updating: participants who scored higher on the KTT had greater differences in their responses for the two conditions, that is, showed a greater adaptation effect. There were no interactions of RMET, ART, or CRT with condition.

Surprisingly, we also found a main effect of ART ($\beta = -106.83$ (52.50), $p < 0.05$), suggesting that participants with greater print exposure provide responses more consistent with the cautious speaker, regardless of condition. However, we believe that this finding should be interpreted with caution: ART was last in our test battery and we had to exclude many participants due to random guesses, suggesting that this measure may not be fully reliable. Additionally, given the lack of an interaction with condition and our focus on how individual differences affect *adaptation*, we leave further investigations of this effect to future work.

The results from this experiment suggest that adaptation to usage patterns of multiple speakers may be modulated by working memory, and, in particular, working memory updating. Since this experiment was purely exploratory, we proceeded by conducting a confirmatory experiment to test whether this result can be replicated. In preparation for the confirmatory experiment, we also estimated another mixed-effects model on all the data available from the adaptation experiment and the KTT task (95 participants after exclusions), which we simplified through backward selection. Also according to this model, we found main effects of condition and most recent speaker, and an interaction of condition and KTT score (see middle column of Table 2).

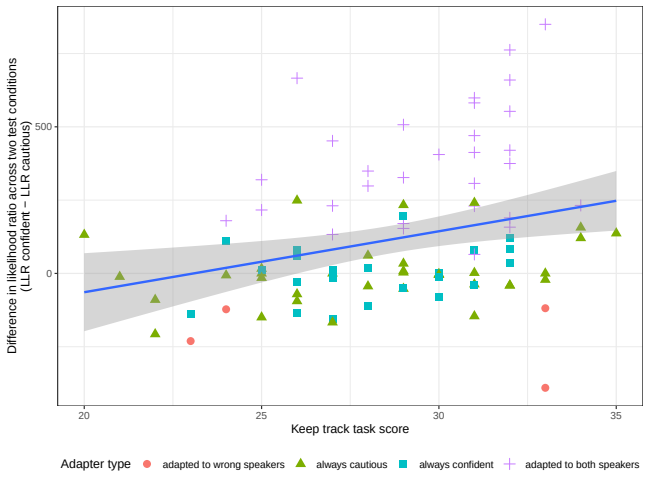

Figure 3: Correlation between difference in log likelihood ratio and performance on the keep track task.

Experiment 2: Confirmatory study

In this experiment, we aimed to replicate the main finding from Exp. 1, i.e., that performance on the keep track task predicts the extent of adaptation. All procedures, exclusion criteria and analyses were preregistered on OSF.[4]

Participants

Based on a power analysis, we recruited 97 native English speakers born and residing in the United States via the crowdsourcing platform Prolific. Participants were paid $6 (∼$12/hr). None of them also participated in Exp. 1.

Procedure

Participants completed the adapatation experiment followed by the KTT.

Results

One participant was excluded based on attention check performance. 3 further participants were excluded for performing more than 2.5 SDs below the mean on the KTT, and 2 additional participants had to be excluded due to missing responses. 91 remaining participants entered further analyses. None of the reported results hinge on these exclusions.

As in the reduced model of Experiment 1, we regressed the log likelihood ratio onto condition, the most recent speaker type, the main effect of KTT and the interaction of KTT with condition. As in the first experiment, we found a significant main effect of condition ($\beta = -58.40$ (11.85), $p < 0.001$) and the effect of which speaker participants saw right before the test blocks ($\beta = -50.16$ (16.02), $p < 0.01$). Crucially, as in Experiment 1, there was no main effect of KTT but a significant interaction of KTT with condition ($\beta = -77.95$ (26.98), $p < 0.01$), suggesting that individuals with greater working memory control adapt more strongly to the two individual speakers. Thus, we replicated the main finding from Exp. 1.

Figure 3 shows the correlation between the KTT score and the difference in log likelihood ratios, where a greater difference indicates that a participant showed stronger speaker-specific adaptation. As in the reanalysis above, we classified participants into four categories, depending on whether their behavior differed across conditions but was the opposite from what was expected (*adapted to wrong speakers*), whether they always provided responses most similar to a "cautious" speaker or always most similar to a "confident" speaker, or whether they *adapted to both speakers*. As this figure shows, most participants who adapted to both speakers, scored above the mean on the KTT task whereas other adapter types had a more even distribution of KTT scores.

General Discussion

In two experiments, we replicated the finding that listeners can adapt to multiple speakers and learn speaker-specific expectations of the use of uncertainty expressions. Furthermore, going beyond these population-level effects, we showed in a re-analysis of the data by S&D that there exists considerable variability in the adaptation behavior across participants, and in the exploratory and confirmatory analyses of how individual differences in cognitive abilities and linguistic experience affect adaptation behavior, we repeatedly found a stable effect of working memory updating on adaptation.

We consider this study an important first step towards a full processing-level account of adaptation. Existing computational accounts based on the ideal observer model make the simplifying assumption that listeners have perfect memory, and S&D even argued against an account according to which listeners struggle with keeping representations of different speakers separate in memory. Our results, on the other hand, indicate that memory limitations play a critical role in semantic-pragmatic adaptation.

An important next question to investigate is the exact link between the KTT and adaptation behavior. One likely explanation of the present results is that participants with higher KTT scores are better at keeping representations of multiple speakers distinct as they are able to maintain a more precise mapping between utterances and speakers, which, in turn, facilitates speaker-specific adaptation. However, given that the keep track task has two components – working memory updating and working memory storage (Ecker et al., 2010; Frischkorn et al., 2022; Panesi, Bandettini, Traverso, & Morra, 2022), it remains unclear which subset of these two components modulates adaptation. Future experiments could test the relationship between adaptation and memory in a dual-task paradigm by manipulating working memory load, as well as by collecting measures of different aspects of working memory. This would help with disentangling the two components of the KTT and their relationship with adaptation, which is an important direction for future work to ultimately inform a full account of adaptation.

[4] https://osf.io/ngack

Acknowledgements

We thank the anonymous reviewers for their thoughtful comments and feedback. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878).

References

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome. *Journal of Child psychology and Psychiatry*, *38*(7), 813–822.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: a new task for assessing theory of mind in adults. *PloS one*, *8*(11), e81279.

Ecker, U. K., Lewandowsky, S., Oberauer, K., & Chee, A. E. (2010). The components of working memory updating: an experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 170.

Fairchild, S., & Papafragou, A. (2021). The role of executive function and theory of mind in pragmatic computations. *Cognitive Science*, *45*(2), e12938.

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, *42*(9), 1362–1376.

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, *11*(5), e0154854.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25–42.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological science*, *17*(2), 172–179.

Frischkorn, G. T., Von Bastian, C. C., Souza, A. S., & Oberauer, K. (2022). Individual differences in updating are not related to reasoning ability and working memory capacity. *Journal of Experimental Psychology: General*, *151*(6), 1341.

Heyman, T., & Schaeken, W. (2015). Some differences in some: examining variability in the interpretation of scalars using latent class analysis. *Psychologica Belgica*, *55*(1), 1.

Johann, V., Könen, T., & Karbach, J. (2020). The unique contribution of working memory, inhibition, cognitive flexibility, and intelligence to reading comprehension and reading speed. *Child Neuropsychology*, *26*(3), 324–344.

Johnson, E., & Arnold, J. E. (2021). Individual differences in print exposure predict use of implicit causality in pronoun comprehension and referential prediction. *Frontiers in Psychology*, 2933.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. doi: 10.1037/a0038695

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Mayn, A., & Demberg, V. (2022). Individual differences in a pragmatic reference game. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

McIlhiney, P., Gignac, G. E., Ecker, U., Kennedy, B., & Weinborn, M. (2022). *Executive function and the continued influence of misinformation: A latent variable analysis.* (PsyArXiv preprint)

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Panesi, S., Bandettini, A., Traverso, L., & Morra, S. (2022). On the relation between the development of working memory updating and working memory capacity in preschoolers. *Journal of Intelligence*, *10*(1), 5.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (irt). *Journal of Behavioral Decision Making*, *29*(5), 453–469.

Roettger, T. B., & Franke, M. (2019). Evidential strength of intonational cues and rational adaptation to (un-)reliable intonation. *Cognitive Science*, *43*(7), e12745. doi: 10.1111/cogs.12745

Scholman, M. C., Demberg, V., & Sanders, T. J. (2020). Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse. *Discourse Processes*, *57*(10), 844–861.

Schuster, S., & Degen, J. (2019). Speaker-specific adaptation to variable use of uncertainty expressions. In *Proceedings of the 41st annual meeting of the Cognitive Science Society (CogSci 2019)*.

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, 104285.

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two-and four-option multiple choice question version of the cognitive reflection test. *Behavior research methods*, *50*(6), 2511–2522.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading research quarterly*, 402–433.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment*

and Decision making, *11*(1), 99.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147–168.

Trott, S., & Bergen, B. (2019). Individual differences in mentalizing capacity predict indirect request comprehension. *Discourse Processes*, *56*(8), 675–707.

Xiang, M., Kramer, A., & Kennedy, C. (2020). *Semantic adaptation in gradable adjective interpretation.* (unpublished MS, U Chicago)

Yang, X., Minai, U., & Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in psychology*, *9*, 1720.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143. doi: 10.1016/j.jml.2015.08.003

Yntema, D. B. (1963). Keeping track of several things at once. *Human Factors*, *5*(1), 7-17. doi: 10.1177/001872086300500102