Quantifying the relationships between linguistic experience, general cognitive skills and linguistic processing skills

Florian Hintz (florian.hintz@mpi.nl)
Max Planck Institute for Psycholinguistics, Wundtlaan 1,
Nijmegen, 6525 XD, The Netherlands

Cesko C. Voeten (CVoeten@fryske-akademy.nl)
Fryske Akademy, Doelestraat 8,
Leeuwarden, 8911 DX, The Netherlands

James M. McQueen (james.mcqueen@donders.ru.nl)
Donders Centre for Cognition, Radboud University, Thomas van Aquinostraat 4,
Nijmegen, 6525 GD, The Netherlands

Antje S. Meyer (antje.meyer@mpi.nl)
Max Planck Institute for Psycholinguistics, Wundtlaan 1,
Nijmegen, 6525 XD, The Netherlands

Abstract

Humans differ greatly in their ability to use language. Contemporary psycholinguistic theories assume that individual differences in language skills arise from variability in linguistic experience and in general cognitive skills. While much previous research has tested the involvement of select verbal and non-verbal variables in select domains of linguistic processing, comprehensive characterizations of the relationships among the skills underlying language use are rare. We contribute to such a research program by re-analyzing a publicly available set of data from 112 young adults tested on 33 behavioral tests. The tests assessed nine key constructs reflecting linguistic processing skills, linguistic experience and general cognitive skills. Correlation and hierarchical clustering analyses of the test scores showed that most of the tests assumed to measure the same construct correlated moderately to strongly and largely clustered together. Furthermore, the results suggest important roles of processing speed in comprehension, and of linguistic experience in production.

Keywords: individual differences; linguistic skills; general cognitive skills

Introduction

Most people acquire their native language effortlessly, yet individuals differ greatly in how they use it (Kidd et al., 2018). The question what makes someone a good language user is intimately connected to that of the cognitive architecture that enables language use (McQueen & Meyer, 2019). Specifically, previous research has established that language processing interfaces with non-verbal cognitive systems (Bates et al., 1991; McClelland & Rumelhart, 1981) rather than being an encapsulated module in the human mind. On such an account, individual differences in linguistic processing skills are in part the result of variability in linguistic experience (leading to variability in knowledge about words and grammatical rules) and variability in general cognitive skills (Dabrowska, 2018; Kidd et al., 2018). While most contemporary theories of language processing are consistent with this view, relevant empirical evidence is still sparse. This is because most studies of individual differences in linguistic processing skills have adopted qualitative approaches (e.g., asking 'is X involved in Y?') and focused on the involvement of just one variable, or a small set of variables in linguistic processing (e.g., working memory in sentence comprehension; but see Christopher et al., 2012; Schmidtke et al., 2018). Such approaches ignore the contribution of other potentially relevant variables and do not allow for a quantification of the relationships between *several* variables. This, however, is critical for understanding how domain-general cognitive skills and linguistic knowledge jointly lead to high or low performance on linguistic tasks.

The present study

In order to quantify the relationships between linguistic knowledge, general cognitive skills and linguistic processing skills, a broad approach is needed where the contributions of multiple predictors are assessed in concert. The goal of the present study was to contribute to such a research program. To that end, we re-analyzed a publicly available dataset containing the scores of 112 participants on 33 behavioral tests (Hintz et al., 2020a). Approximately half of these tests assessed participants' linguistic processing skills on word- and sentence-level production and comprehension tasks. The other tests assessed five cognitive constructs that have been implicated in word and sentence production and/or comprehension in earlier work. These were (1) linguistic experience, i.e. knowledge of words and grammar, (2) processing speed, (3) working memory, (4) inhibition, and (5) non-verbal intelligence. With the exception of (5) each skill was assessed in multiple tests, thereby alleviating concerns about task impurity (Miyake et al., 2000). The importance of each of the skills has been amply demonstrated in earlier work (for (1), see, for instance, Diependaele et al., 2013; Mainz et al., 2017; Huettig & Pickering, 2019); for (2) Engelhardt et al., 2018; Hintz et al., 2020b; Schubert et al.,

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

2015, 2017; for (3) Just & Carpenter, 1992; Baddeley, 2003; Martin, 2021; for (4) Shao et al., 2012; for (5) Engelhardt et al., 2018; Deary, 2001; Visser et al., 2006). However, no study has analyzed the joint impact of all of these variables on word or sentence production or comprehension.

Hintz et al. (2020a) provide pre-processed (e.g., outlier-excluded) scores of tests tapping into word- and sentence-level production and comprehension, and the skills mentioned above (see Table 1 for an overview). Here, we quantified the relationships between these test scores by applying correlation (presented as a heatmap) and hierarchical clustering analyses (HCAs). HCA is a statistical method, which – similar to correlation testing – aims to assess the similarity of scores (i.e., clusters). Going beyond correlation analysis, HCA builds a hierarchy of clusters. To that end, on the agglomerative approach used here, each score is initially assigned to its own cluster. The algorithm then proceeds iteratively, such that at each stage the two most similar clusters are joined, continuing until there is just a single cluster left. HCA does not require any *a priori* choices of test groupings and is therefore a suitable tool for an unbiased exploration of the quantitative relationships between test scores. The most common output of an HCA is a dendogram—a diagram representing a hierarchically-structured tree. Closeness and distance of scores in the tree reflect greater and lesser amounts of shared variance, respectively.

Our analyses were exploratory. We expected that scores of tests assumed to measure the same cognitive construct would show moderate to strong correlations (reflected in warm colors in the heatmap) and would cluster together in the dendogram. Beyond that, no specific hypotheses were formulated. The main issue to be explored was how strongly the scores from the word- and sentence-level production and comprehension tasks were related to each other and to the scores reflecting linguistic experience, general intelligence, processing speed, working memory, and inhibition**.**

Methods

Participants

The dataset provided by Hintz et al. (2020a) contains data from 112 participants. Eighty-seven of these were university students or graduates (27 male, mean age: 22.6 years, range: 18 to 29 years); 24 attended or had attended a vocational college (12 male, mean age: 21.0 years, range: 18 to 29 years), and one participant was a high school graduate (female, 25 years). All participants were native speakers of Dutch. Although higher numbers of participants would be desirable for an individual-differences study, 112 participants were sufficient to detect correlations of .3 with 80% power. Note that the same power calculation applies to the HCA since HCA relies on a distance matrix, which in our study was just a linear transformation of the correlation matrix for our tests into the corresponding Euclidean distances. Nevertheless, we assessed the stability of the dendogram by means of a permutation test. That is, we randomly permuted (n = 1,000) the original distance matrix and each time built a dendogram, comparing the predicted classes of the original dendrogram to the dendogram based on the randomly-permuted distance matrix. The mean absolute correlation of the original dendrogram to the 1,000 permuted dendrograms was .03. In other words, there was a 3% chance of obtaining the dendogram below if the clustering of the tests in Hintz et al. (2020a) were random. We thus conclude that our HCA was sufficiently stable given these data.

Materials and procedure

The 33 behavioral tests administered by Hintz and colleagues (2020a) required participants to provide speeded and unspeeded manual (e.g. button presses or mouse clicks) and spoken responses (e.g., picture naming). Administration took approximately four hours per participant and was divided into four sessions of one hour each (two in the morning and two in the afternoon of the same day). The tests and the items within each test were presented in a fixed order. To assess the tests' retest reliability, all participants completed the same test protocol a second time, approximately four weeks after the first test day. For a detailed description of the test protocol, the individual test materials and procedures, see Hintz et al. (2020a). Table 1 provides an overview of the tests, the cognitive construct that each test was assumed to measure, the nature of the dependent variables (accuracy- vs. RT-based) and key descriptive statistics. Internal consistency and test-retest reliability was excellent for most tests, with the exception of the Corsi block and Flanker tests. This was unexpected, as these are standard frequently used and well validated tests. The authors noted that in both tests participants were excluded based on task misunderstandings. One possibility is thus that poor internal consistency and retest reliability are the result of technical issues in the test administration. Moreover, descriptive statistics revealed poor results for the newly developed monitoring in (non-)word lists and sentences-in-noise tests, which – as the authors explain – might have been too difficult for the participants. This means that no strong conclusions should be drawn from results involving those tests.

Data analysis

We harmonized the scores such that for all tests higher scores reflected better performance. No pre-processing was applied to the data other than that done by Hintz et al. (2020a). The dataset had some missing values, which were missing at random. Over the 112 participants, 0.46 values were missing on average (min = 0, max = 5); over the 35 scores, 1.46 values were missing on average (min = 0, max = 7). This amounted to a total of 1.30% of missing data. We imputed missing data points using MICE (van Buuren & Groothuis-Oudshoorn, 2011), as implemented in the function

'mice' from the eponymous R package. The resulting data matrix was submitted to a correlation analysis. We additionally transformed the correlation matrix into a Euclidean distance matrix, using the function 'cor2dist' from R package psych. The Euclidian distance matrix served as input to the HCA, which was performed using R function hclust using the complete-linkage method.

Table 1: Overview of the tests provided by Hintz et al. (2020a), the cognitive constructs measured and key descriptive statistics for each dependent variable.

Construct measured	Test	DV	N	Skew.	Kurt.	Intern. consist.	Retest reliab.
Linguistic knowledge	Peabody	Acc.	112	-0.43	-0.79	0.96	0.91
Linguistic knowledge	Antonym production	Acc.	111	-0.28	-0.34	0.70	0.74
Linguistic knowledge	Spelling	Acc.	112	-0.43	-0.37	0.83	0.85
Linguistic knowledge	Author recognition	Acc.	112	0.62	0.47	0.93	0.95
Linguistic knowledge	Idioms	Acc.	112	-0.33	0.03	0.53	0.78
Linguistic knowledge	Prescriptive grammar	Acc.	112	0.04	-0.65	0.74	0.86
Processing speed	Auditory simple RT	RT	112	-1.36	3.10	0.90	0.59
Processing speed	Auditory choice RT	RT	112	-0.60	0.15	0.96	0.76
Processing speed	Letter comparison	RT	107	0.65	0.47	0.89	0.83
Processing speed	Visual simple RT	RT	112	-0.54	0.51	0.86	0.58
Processing speed	Visual choice RT	RT	112	0.88	0.73	0.95	0.78
Working memory	Digit span forward	Acc.	112	0.26	-0.74	0.81	0.75
Working memory	Digit span backward	Acc.	110	0.22	-0.56	0.72	0.70
Working memory	Corsi forward	Acc.	111	-0.08	0.25	0.53	0.39
Working memory	Corsi backward	Acc.	108	-0.04	-0.15	0.71	0.49
Inhibition	Flanker	RT	106	-0.57	2.77	0.98	0.50
Inhibition	Antisaccade	Acc.	111	-2.09	6.40	0.89	0.71
Non-verbal IQ	Ravens	Acc.	112	-0.48	-0.33	0.87	0.87
Word production	Picture naming	RT	111	-0.28	0.76	0.88	0.69
Word production	Verbal fluency (Semantic categories)	Acc.	106	0.04	-0.18	-	0.72
Word production	Verbal fluency (Letters)	Acc.	112	0.15	0.55	-	0.71
Word production	Maximal speech rate	RT	106	-0.24	-0.07	-	0.88
Word production	Word pronunciation	Acc.	111	-0.12	-0.57	0.46	0.79
Word production	Non-word pronunciation	Acc.	111	0.26	0.60	0.88	0.88
Word comprehension	Non−word monitoring in noise	Acc.	112	-2.11	7.87	-	0.59
Word comprehension	Word form monitoring in noise	Acc.	112	-2.47	9.98	-	0.49
Word comprehension	Meaning monitoring in noise	Acc.	109	-1.02	2.48	-	0.53
Word comprehension	Rhyme judgment	RT	109	-0.49	0.10	0.94	0.79
Word comprehension	Lexical decision	RT	112	0.61	0.72	0.97	0.69
Word comprehension	Semantic categorization	RT	109	-0.90	0.72	0.96	0.62
Sentence production	Phrase generation	RT	112	0.47	0.80	0.82	0.79
Sentence production	Sentence generation	Acc.	112	-1.14	0.98	0.95	0.67
Sentence comprehension	Gender prediction	RT	105	0.45	-0.95	0.88	0.88
Sentence comprehension	Verb prediction	RT	112	0.62	-0.72	0.86	0.76
Sentence comprehension	Sentence monitoring in noise	Acc.	112	-0.67	0.16	-	0.30

Note. DV = dependent variable, Skew. = skewness, Kurt. = kurtosis, Intern. consist. = internal consistency, Retest reliab. = retest reliability

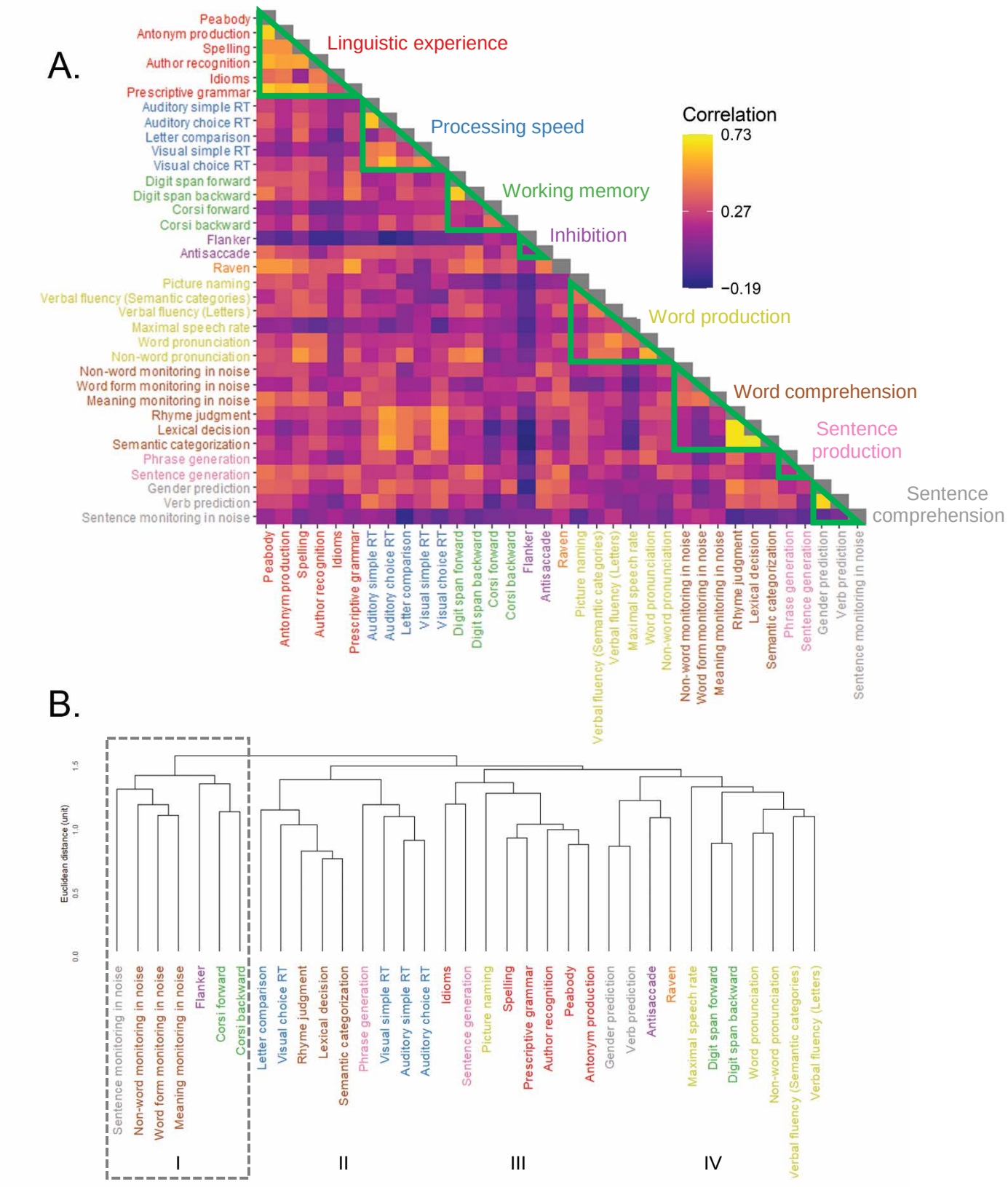

Figure 1: Relationships between test scores. Panel A: Correlation matrix presented as a heatmap. Warm colors represent strong correlations. Scale ranges from weakest to strongest correlation observed. Correlation coefficients from the same cognitive construct enclosed in green triangles. Panel B: Dendogram. Statistically similar scores form clusters.

Results and discussion

Figure 1 summarizes the results of the correlation analysis (Panel A) and the HCA (Panel B). Correlations between test scores ranged between -.19 and .73. As expected, tests assumed to measure the same cognitive construct generally correlated positively with each other, moderately to strongly, as reflected in warm-color cells (e.g., orange-yellow) enclosed within the green triangles along the gray diagonal in the heatmap. In addition to the tests with poor descriptive statistics mentioned above (Corsi, Flanker, (non-)word and sentence monitoring), which correlated only weakly with other tests assumed to measure the same cognitive construct, we observed that the letter comparison test correlated less strongly with the other processing speed tests, that picture naming was correlated less strongly with the other word production tests, and that phrase and sentence generation tests were not strongly correlated.

Turning to the hierarchical representation of the relationships between the test scores, in the dendogram, the first split divided the tree into two main branches: One branch consisted of three sub-clusters (II, III, and IV); the other branch (sub-cluster I) featured seven test scores that were grouped together by the algorithm but were completely unrelated to the remaining tests. These latter tests, i.e. monitoring in noise, Flanker, and Corsi Block tests, were thus unrelated to other tests assumed to measure the same cognitive construct (i.e., word and sentence comprehension, inhibition and working memory) and did not show relationships with tests measuring other cognitive constructs either. The most likely explanation for this separation is – as mentioned earlier – the tests' poor performance on internal consistency and test-retest reliability. This sub-cluster of tests is not considered further here.

As expected and also shown in the heatmap, the tests assumed to tap a common construct generally appeared close together in the cluster structure. That is, the tests of processing speed occurred together in cluster II, the remaining domain-general skills in cluster IV, and the linguistic experience tests in cluster III.

More interestingly, the word and sentence production tasks did not appear in close proximity in the dendogram, but were distributed over different sub-clusters. This reflects that they depended to differing degrees on specific domain-general skills and linguistic experience. Specifically, the phrase generation task, where participants were familiarized with a small set of common objects, which they subsequently named in noun and adjectival phrases of increasing complexity, was strongly affiliated with the processing speed tasks. By contrast, sentence generation, where participants produced transitive sentences in active and passive voice, and picture naming, where they had to retrieve the high and low frequency names of pictures, were closest to the linguistic knowledge tasks. This is plausible (at least with the wisdom of hindsight), because the phrase generation task only required participants to rapidly order a small set of known words, whereas picture naming required the retrieval of lexical knowledge. This clustering is consistent with the entrenchment hypothesis (e.g., Diependaele et al., 2013), according to which linguistic experience facilitates lexical access. The remaining word production tasks, including the verbal fluency tasks, clustered together with the tests of working memory, again (for verbal fluency tasks) consistent with earlier work (Shao et al., 2011).

Turning to the comprehension test scores, the speeded word comprehension tasks (i.e., lexical decision, semantic categorization, rhyme judgment) clustered together with the processing speed tasks, whereas the sentence comprehension tasks, where participants had to predict upcoming words based on gender cues and verb semantics, clustered together with other domain-general processing tasks, i.e. the Raven non-verbal intelligence test, the digit span working memory tests and the antisaccade test, as well as the majority of word production tests. The presence of the antisaccade test in that cluster is unexpected and not straightforwardly motivated by prior research or models of sentence comprehension. Mediating influences of working memory on visually-aided auditory prediction tasks, on the other hand, have previously been reported (Huettig & Janse, 2016). Moreover, prominent theories have linked prediction during comprehension to word production (Dell & Chang, 2014; Pickering & Garrod, 2007; see Hintz et al., 2017, and Rommers et al., 2015, for experimental evidence for a link between prediction and production abilities).

The scattering of the word and sentence production and comprehension tasks over different clusters shows, first, that they are not as closely related as one might have thought (cf. Chater et al., 2016), and second, that they draw to differing degrees on domain-general skills and linguistic knowledge. Given the nature and complexity of the tasks, this not surprising. In evaluating the present findings, it is important to keep in mind that linguistic processing, in particular sentence comprehension, was only assessed by a small set of tasks. Therefore, general conclusions about the relationships between the production and comprehension system and their reliance on linguistic knowledge and domain-general skills can only be drawn on the basis of further work. The present study illustrates how such work could proceed in the future, and what can be gained by individual-differences studies that simultaneously assess multiple skills in each participant.

Conclusion

The present re-analysis of the public dataset by Hintz et al. (2020a) was exploratory. We expected that scores of tests assumed to measure the same cognitive construct would cluster together. The main issue to be explored was how strongly the scores from the production and comprehension tasks were related to each other and to the scores reflecting linguistic experience, non-verbal intelligence, processing speed, working memory and inhibition**.** We observed a strong association between processing speed and word comprehension, and between linguistic experience and production. Interestingly, we saw separate sub-clusters for comprehension and production, suggesting that these

processes are less strongly related than one might have thought.

References

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*(3), 189-208.

Bates, E., Bretherton, I., & Snyder, L. S. (1991). *From first words to grammar: Individual differences and dissociable mechanisms* (Vol. 20). Cambridge University Press.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45(3),* 1-67. Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language, 89*, 244-254.

Christopher, M. E., Miyake, A., Keenan, J. M., Pennington, B., DeFries, J. C., Wadsworth, S. J., ... & Olson, R. K. (2012). Predicting word reading and comprehension with executive function and speed measures across development: a latent variable analysis. *Journal of Experimental Psychology: General*, *141*(3), 470-488.

Dąbrowska, E. (2018). Experience, aptitude and individual differences in native language ultimate attainment. *Cognition, 178*, 222-235.

Deary, I. J. (2001). Human intelligence differences: A recent history. *Trends in Cognitive Sciences, 5*, 127–130.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*: 20120394.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first-and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology, 66*(5), 843-863.

Engelhardt, P. E., Nigg, J. T., & Ferreira, F. (2017). Executive function and intelligence in the resolution of temporary syntactic ambiguity: an individual differences investigation. *Quarterly Journal of Experimental Psychology, 70*(7), 1263-1281.

Hintz, F., Dijkhuis, M., Van 't Hoff, V., McQueen, J. M., & Meyer, A. S. (2020a). A behavioural dataset for studying individual differences in language skills. *Scientific Data, 7*: 429.

Hintz, F., Jongman, S. R., Dijkhuis, M., Van 't Hoff, V., McQueen, J. M., & Meyer, A. S. (2020b). Shared lexical access processes in speaking and listening? An individual differences study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(6), 1048-1063.

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(9), 1352-1374.

Huettig, F., & Pickering, M. J. (2019). Literacy advantages beyond reading: Prediction of spoken language. Trends in *Cognitive Sciences, 23*(6), 464-475.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review, 99*(1), 122-149.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences, 22*(2), 154-169.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375.

Mainz, N., Shao, Z., Brysbaert, M., & Meyer, A. S. (2017). Vocabulary knowledge predicts lexical processing: Evidence from a group of participants with diverse educational backgrounds. *Frontiers in Psychology, 8:* 1164.

Martin, R. C. (2021). The critical role of semantic working memory in language comprehension and production. *Current Directions in Psychological Science, 30*(4), 283-291.

McQueen, J. M., & Meyer, A. S. (2019). Key issues and future directions: Towards a comprehensive cognitive architecture for language use. In P. Hagoort (Ed.), *Human language: From genes and brain to behavior* (pp. 85-96). Cambridge, MA: MIT Press.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension?. *Trends in Cognitive Sciences, 11*(3), 105-110.

Rommers, J., Meyer, A. S., & Huettig, F. (2015). Verbal and nonverbal predictors of language-mediated anticipatory eye movements. *Attention, Perception & Psychophysics, 77*(3), 720-730.

Schubert, A. L., Hagemann, D., Voss, A., Schankin, A., & Bergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence, 51*, 28-46.

Schubert, A. L., Hagemann, D., & Frischkorn, G. T. (2017). Is general intelligence little more than the speed of higher-order processing?. *Journal of Experimental Psychology: General, 146*(10), 1498-1512.

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2018). Individual variability in the semantic processing of English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(3), 421-439.

Shao, Z., Roelofs, A., & Meyer, A. S. (2012). Sources of individual differences in the speed of naming objects and actions: The contribution of executive control. *Quarterly Journal of Experimental Psychology, 65*(10), 1927-1944.

Visser, B. A., Ashton, M. C., & Vernon, P. A. (2006). Beyond g: Putting multiple intelligences theory to the test. *Intelligence, 34*(5), 487-502.