A comparison of meta-analysis, mega-analysis, and a hybrid approach

Ezequiel Koile[a], Sho Tsuji[b], & Alejandrina Cristia[c]

[a] ADD

[b] ADD

[c] ADD

Author Note

Correspondence concerning this article should be addressed to Ezequiel Koile, ADD.
E-mail: ADD

Abstract

Laboratory measures of infant speech perception have been central to the development of theories of infant language acquisition, and could be valuable predictors of important individual and group variation. A recent report suggests that these measures' psychometric properties may be limited, based on a meta-analytic analysis. We re-analyze those data using a mega-analytic approach, as well as a variety of hybrid approaches. We find that (a) the results of meta- and mega-analyses diverge significantly, and (b) a mega-analytic approach can be more powerful in detecting stability in performance across days. However, since it is often difficult to recover original data, we also explore a hybrid approach, in which some studies are represented by group statistics, and others by the original data, assessing to what extent biased data sharing may impact overall conclusions.

A comparison of meta-analysis, mega-analysis, and a hybrid approach

Recent years have seen the rise of cumulative science, in which each new result is integrated into the web of prior knowledge. In this paper, we introduce mega-analyses, a cumulative science method that is rare in the study of cognition. Mega-analyses involve integrated analyses of raw data collected in multiple sites using a single pre-processing and statistical analysis pipeline. They thus differ from simple analyses in the scope of the data, dealing with more heterogeneous sets since the sites may not have collected data in a coordinated manner; and from meta-analyses in that the raw data are inputed, rather than group-based statistics. We couch this presentation in the context of one case study to facilitate a discussion of in which contexts is a meta- or mega-analyses more appropriate tool in the context of cumulative science.

Study case: Reliability of infant speech perception measures

Infant speech perception measures have been central to the development of theories of language acquisition. For example, experimental measures showing that infants' perception for non-native contrasts varied between 6 and 12 months of age led to the conclusion that phonological acquisition begins as early as this (Werker & Tees, 1984). More recently, these same measures have been argued to be valuable predictors of meaningful individual and group variation. Cristia, Seidl, Junge, Soderstrom, and Hagoort (2014) meta-analyzed 20 articles and theses reporting correlations between speech perception measures (including, for example, perception of non-native contrasts) and vocabulary; as well as work comparing performance in such tasks by infants at risk of a language disorder against infants not at risk. The authors concluded that individual and group variation was significantly associated with performance in infant speech perception tasks (median r=.31, 95% CI [.22, .40]), in line with the hypothesis that infant speech perception measures can provide an insight into individual infants' language skills.

One outstanding issue, however, concerns the psychometric properties of such measures, and in particular their reliability. We will call the correlation between two versions of a given measure (such as a test-retest correlation) its *reliability*; and the correlation between that measure and a measure of something else (a potential predictor or predicted variable) its *validity*. Demonstrations within classical test theory suggest that the validity of any measure is bounded by the square root of its reliability (e.g., Michell, 2003).

Only two studies have been published reporting on test-retest correlations of infants undergoing the same speech perception measures twice (Cristia, Seidl, Singh, & Houston, 2016; Houston, Horn, Qi, Ting, & Gao, 2007). Since the later paper contains data on the first, we only discuss the later on. Cristia et al. (2016) used a meta-analytic method to combine the earlier results with test-retest data collected independently by three research labs (each of which carried out 3-5 studies), which did not know the others were also gathering the same kind of data. Meta-analytic methods seem to conceptually fit well the goal of integrating results in such a setting. Thus, the authors first estimated the test-retest correlation for each of the 12 studies (13 with Houston et al., 2007), and then derived the overall estimate as the weighted median correlation. Surprisingly, this revealed an overall r=0.065, with a 95% confidence interval of [-0.12; 0.25]. This results was not just due to some of the studies providing very small correlation coefficients, but crucially because some of these coefficients were negative.

If these results are to be believed, this means that correlation work reporting on these measures' validity (e.g., correlations with vocabulary estimates, group differences) is suspect, because the measures' null reliability would entail that no validity can be measured. Cristia et al. (2016) made the case that it was appropriate to integrate across all 13 studies because there was no reason to believe that test-retest would yield negative correlations. While this is true, calculating correlations as a measure of test-retest stability within each study and then averaging them is not equivalent to calculating test-retest stability in behavior taking all studies into account together.

In fact, genetics and neuroimaging research are seeing the emergence of work that discusses the benefits of considering raw data together in what are called mega-analyses [e.g., CITATIONS HERE]. Mega-analyses are generally preferrable over meta-analyses because pre-processing steps can be done in a homogeneous fashion, removing this potential source of variance. As it happens, Cristia et al. (2016) pre-processed all data (except for the published study) in the same way, and thus this was not a consideration. Additionally, a second and crucial advantage of mega- over meta-analyses is that structured sources of variance can be better accounted for, and analyses therefore have more power to detect small and stable effects.

The present study

We reanalyze Cristia et al. (2016)'s data to revisit the question of how reliable infant speech perception tasks are using a mega-analytic approach. We address this question first assuming that the analyzer has access to all data, which is the case here. This answer is most informative for readers that are specifically interested in the question of reliability. We then make a few reasonable assumptions regarding data missingness. If starting a mega-analysis from scratch, the cumulative scientist may only have access to a subset of the raw data, with the remaining studies in the literature being represented by group-based summary statistics at best. Althow raw data may be missing at chance, we also contemplate three cases of biased missingness that would be due to selective reporting. First, we assume data would be missing for studies with a small number of participants; second for studies with small main effects; and third for those with small or negative test-retest correlations (assuming that the original researchers finding themselves in one of these situations may be less prone to make the additional effort of sharing the raw data).

Methods

Very short because we refer to previous paper for full description of experiments

table of experiments: short names, short description, N of children, mean age

We got data from osf, using R, this paper uses Rmd in RStudio & papaja for increased reproducibility.

Results

- how should structure be accounted for - are studies all different from each other?

explain use of AIC to compare models, using also conceptual reasons to group studies – ending up with 5 clusters

- in mega-analysis, do you also find basically no prediction of test2 from test1?

  no, we get something pretty different. Explain why

- what happens if you only have some data from some studies – picked at random? (assuming original authors do not withhold the data for any reason that is related to the data itself)

- and if you only have data from large studies? (authors who ran more babies are more motivated to share)

- and if you only have data from studies with large main effects? (defined as the average between effect at test1 and effect at test2 – intuition is that authors with strong effects believe their data more)

- and if you only have data from studies with large test-retest correlations? (idea: authors who find reliability more likely to share raw data)

- use a graph to represent all of the hybrid results (max 4k words!)

Discussion

- under what conditions can we trust infant speech perception measures of individual variation?
- we recommend mega- over meta-analysis
- explain under what conditions this holds, and when mega-analysis provides biased view of data

References

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test-retest reliability in infant speech perception tasks. *Infancy*, *21*(5), 648–667. Retrieved from https://osf.io/62nrk/

Houston, D., Horn, D. L., Qi, R., Ting, J. Y., & Gao, S. (2007). Assessing speech discrimination in individual infants. *Infancy*, *12*, 119–145.

Michell, J. (2003). Measurement: A beginner's guide. *Journal of Applied Measurement*, *4*(4), 298–308.

Werker, J. F., & Tees, R. (1984). Cross-language speech perception. *Infant Behavior and Development*, *7*, 49–63.