Attentional Control Data Collection: A Resource for Efficient Data Reuse

Julia M. Haaf[1], Madlen Hoffstadt[1], & Sven Lesche[2]

[1] University of Amsterdam

[2] University of Heidelberg

Version 1, October 2023

Author Note

Correspondence concerning this article should be addressed to Julia M. Haaf, Nieuwe Achtergracht 129B, 1018 WT Amsterdam, The Netherlands.. E-mail: j.m.haaf@uva.nl

17   Abstract

18   One or two sentences providing a **basic introduction** to the field, comprehensible to a

19   scientist in any discipline.

20   Two to three sentences of **more detailed background**, comprehensible to scientists

21   in related disciplines.

22   One sentence clearly stating the **general problem** being addressed by this particular

23   study.

24   One sentence summarizing the main result (with the words "**here we show**" or their

25   equivalent).

26   Two or three sentences explaining what the **main result** reveals in direct comparison

27   to what was thought to be the case previously, or how the main result adds to previous

28   knowledge.

29   One or two sentences to put the results into a more **general context**.

30   Two or three sentences to provide a **broader perspective**, readily comprehensible to

31   a scientist in any discipline.

32   *Keywords:* Open Data, Attentional Control, SQL

33   Word count: X

Attentional Control Data Collection: A Resource for Efficient Data Reuse

Making data openly available has been a central demand by reformers since the start of the reproducibility crisis in psychology [REFS]. Fortunately, this demand has lead to a considerable increase in data availability. While only about 25% of data were shared after request in 2006 (Wicherts, Borsboom, Kats, & Molenaar, 2006), publicly sharing data upon publication is now more and more the norm. This cultural shift is also increasingly institutionalized. Universities and funding agencies prioritize open data, and some journals even mandate the publication of data with every published article (Sloman, 2015). In addition, technology like the Open Science Framework (OSF) and other data sharing facilities enable an easy process for researchers, further reducing barriers to share data.

Data sharing serves two goals: 1. To make the scientific process more transparent and enable error and fraud detection, and 2. to make the scientific process more efficient by allowing data reuse for different research projects. Current data sharing efforts, however, seemingly focus overwhelmingly on the first goal [REF Cruewell et al, 2023]. Whenever researchers complying with common data sharing procedures publish an article, they share the corresponding data on the OSF, ideally in a format that allows to redo the exact analyses reported in the article. The OSF repository is linked in the article, and readers may access the data through this link and check whether analysis code and shared data correspond to the results section in the article. This setup, while appropriate for the first goal of data sharing, ignores the second goal of data reuse.

To enable data reuse, data sharing needs to be approached differently. For example, consider a researcher (like the first author of the current paper) might me interested in the Stroop task (Stroop, 1935). The Stroop task is popular in cognitive psychology (MacLeod, 1991), so we may assume that many studies include this or similar tasks in their studies. Instead of running yet another Stroop experiment, the researcher decides to use existing data to explore their research question before designing a more targeted study. First, the

60 researcher needs to be able to find open Stroop task data. Currently, they could either

61 search for papers on the topic and check whether open data are provided, or search directly

62 via OSF or other data sharing servers. However, neither of these options is very promising as

63 the vast majority of articles in the literature does not provide raw data and data sharing

64 servers are not equipped with sufficient search options. Second, data sets need to be accessed

65 easily and in a general, understandable format ready for reuse. There are data sharing

66 formats that provide this structure [REF], but they are rarely used. Additionally, data are

67 usually shared on the level necessary for the original analysis. In case of the Stroop task,

68 shared data might provide the Stroop effect per participant, but for this new analysis the

69 researcher needs trial-level data. So again, there is yet another barrier for data reuse.

70 We think it is necessary to provide a data sharing solution that solves the current

71 problems and enables easy and efficient data reuse. Here, we propose to gather open data

72 sets from a specific research area in an SQL data base. This process requires little to no

73 work in addition to current data sharing policies from the authors of original papers, some

74 work from the lab(s) setting up the data base, and little to no work from the researchers who

75 wish to reuse open data. We describe the process and structure we used to set up a data

76 base of attentional control tasks called the Attentional Control Data Collection (ACDC).

77 The data base includes XXX data sets from XXX publications from tasks like the Stroop,

78 Simon, and flanker tasks. Subsequently, we show how the data can be explored using a Shiny

79 app and accessed using an R-package. In an example analysis, we assess the reliability of the

80 included tasks. This section highlights how an open data base like ACDC can aid

81 meta-analytic efforts as well as methodological innovation.

82 To provide a little history of the project, the Attentional Control Data Collection was

83 inspired by a collection of open data sets from attentional control tasks by the Perception

84 and Cognition Lab led by J. Rouder (url). Colleagues provided the first author and Rouder

85 with data sets for their statistical work (Haaf & Rouder, 2017; Rouder, Kumar, & Haaf,

2023). To ensure that data sets were accessible, we gathered them in a github repository. However, there was little structure to the collection, and github repositories are neither stable entities nor are they designed as data storage. Here, we describe how a structured data collection can be achieved and which benefits it provides.

SQLight Database

One of the most standard ways in computer science for storing data is using a Structured query language (SQL) data base. SQL allows to create, access and manipulate a structured data storage. SQL data bases consist of data tables and relations between these tables. There are many flavors of SQL data bases. Here, we decided to use an SQLight data base, a lightweight solution that allows us to store the entire data base in a single file of moderate size that can be downloaded by researchers for data reuse. In this section we describe the structure of the data base and the data currently included. Researchers who simply want to use ACDC may safely skip this section.

Database Structure

SQL databases are composed of several data tables consisting of rows and columns. Each row in a data table has a primary key (essentially a row ID) which uniquely identifies it. To connect one table to another table in the database, tables may contain foreign keys which reference a unique row in another data table. In contrast to primary keys, these foreign keys can have duplicate values in the data table, as long as they are unique in their primary table. For instance, a study table may store information about all studies in a database where each row corresponds to a single study. Here the primary key is the study_id. We can ensure that our database links each study to the publication it was published in by adding a foreign key called publication_id. This foreign key references the unique identifier of the respective publication in a publication table. Figure 1 illustrates the relationship between study table and publication table. While publication_id links to a single row in the publication table, it can occur several times in the study_table as there can be several studies per publication.
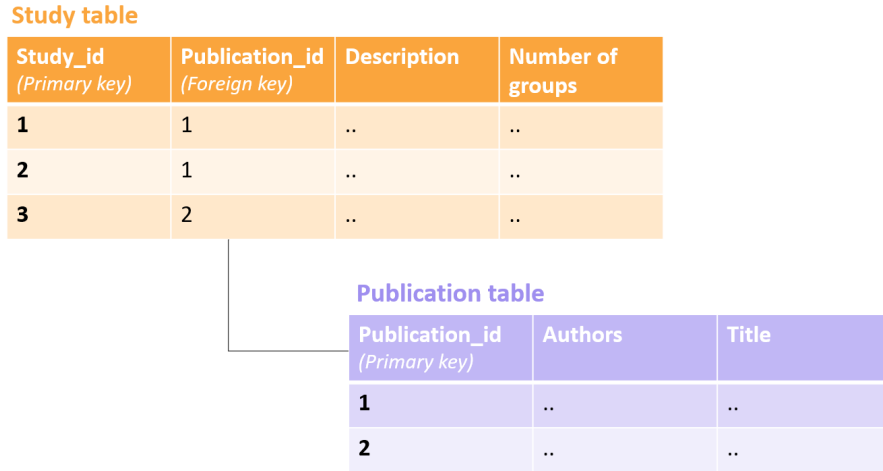
Study table

Study_id (Primary key)	Publication_id (Foreign key)	Description	Number of groups
1	1	..	..
2	1	..	..
3	2	..	..

Publication table

Publication_id (Primary key)	Authors	Title
1	..	..
2	..	..

Figure 1. Illustrative example of using foreign and primary keys in a SQL database.

¹¹² ACDC is adapted to the logic of publications consisting of one or multiple studies

¹¹³ which in turn include one or several data sets. The whole structure of ACDC is shown in

¹¹⁴ Figure 2. A *publication table* and a *study table* contain specific information about each

¹¹⁵ publication and study, respectively. Each data set within a study stores trial-level

¹¹⁶ information about a single attentional control task within a certain study. If a

¹¹⁷ between-subject manipulation exists, our data base contains a separate data set for each

¹¹⁸ group and each task. For instance, a study in which two groups (younger and older adults)

¹¹⁹ who completed a Simon and a Stroop task would consist of four data sets (i.e.,

¹²⁰ younger-Stroop, younger-Simon, older-Stroop, older-Simon) in the ACDC data base.

¹²¹ A *data set table* stores information about each data set, such as sample size, the number

¹²² of within-participant manipulations, and whether a fixation cross was used. The *observation*

¹²³ *table* hold the trial-level attentional control task data (including response time and accuracy).

¹²⁴ The task type of each data set (i.e., Stroop, Simon, Flanker, negative priming, or other) and

¹²⁵ a description of which stimuli were presented in the task are documented in the *task table.*

¹²⁶ Note that since the congruency of stimuli, i.e., whether response and stimulus

¹²⁷ attributes are compatible or incompatible, is part of every attentional control task, it is not

considered a separate within-participant manipulation in this database but is per default included in the observation table. Any additional within-participant manipulations are coded in the within ID column of the observation table. Further information about each condition of each data set (such as the percentage of congruent trials, mean response time, and mean accuracy) are recorded in the *within table*.
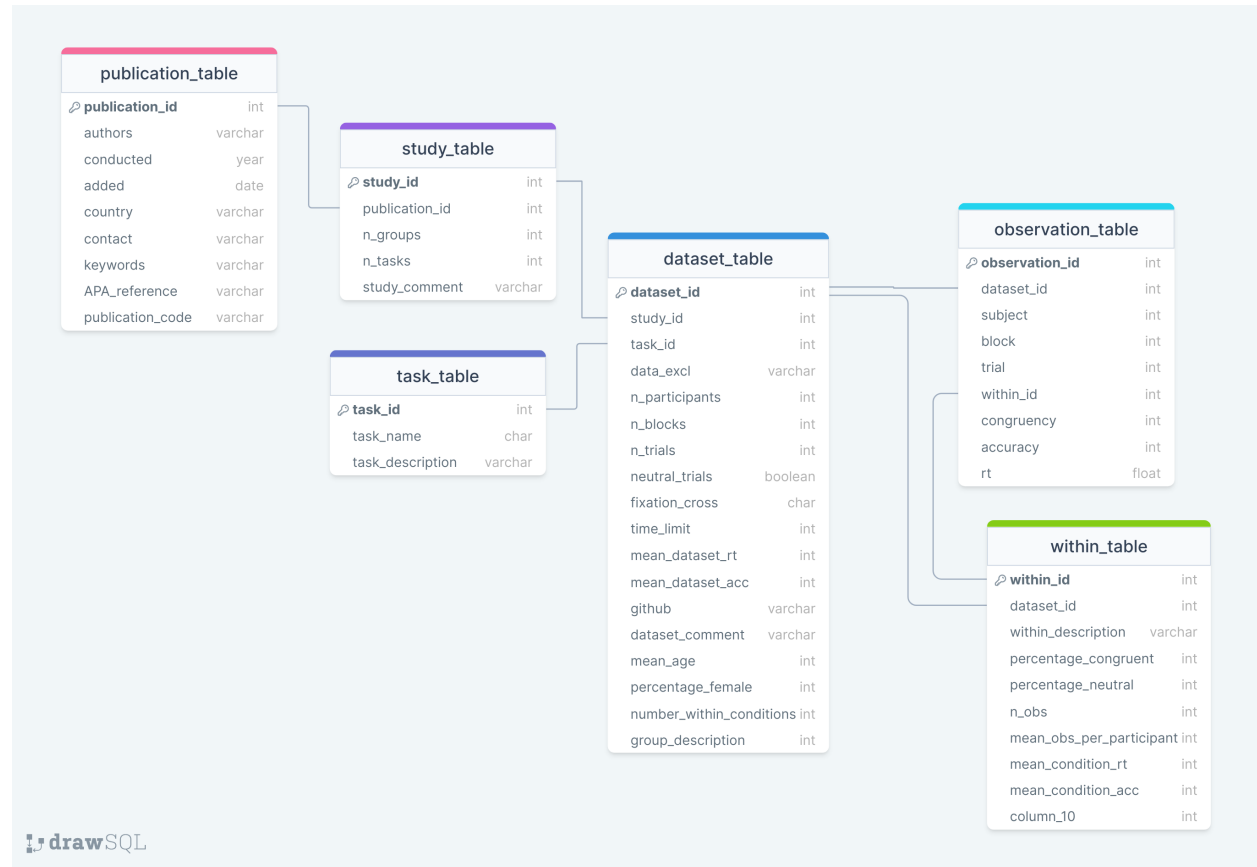

Figure 2. Structure of the ACDC database. Primary keys are indicated by the key symbol. References between data tables are illustrated through lines connecting columns across data tables. This overview includes the data type of each column: integers (int), numbers with decimal places (float), characters (varchar) and logical true/false values (Booleans).

Included Data

Up to the date of submission of the manuscript, 54 data sets from 11 publications are included in the database. The full list of data sets and references is provided in the Appendix. The current database includes data sets from studies with an experimental as well as a correlational focus. The data contain 10^6 observations collected from 7486 participants.

¹³⁸ We did not conduct a systematic search for data sets nor attempted to distribute a
¹³⁹ wider call for open data. Instead, we included the data sets that were already made available
¹⁴⁰ to the lab for previous projects, and added data sets from collaborators bit by bit. This
¹⁴¹ approach was chosen to make the project feasible, and to first set up a working data base
¹⁴² before large quantities of data are added.

Contribute
¹⁴³

Julia ¹⁴⁴ *(TODO: Add a brief description on how researcher can add their own data.)*

¹⁴⁵ <div align="center">**Accessing the Database**</div>

¹⁴⁶ One advantage of SQLite databases is that they are simply a file that can be
¹⁴⁷ downloaded and locally accesses by anyone. Our database is provided in a github
¹⁴⁸ repository.[1] To access the database, researchers can download the file `acdc.db`, and use the
¹⁴⁹ SQLite tool of their choice. In addition, we build `R`-based tools to inspect, access, and select
¹⁵⁰ data from ACDC. We introduce these tools, a shiny app and an `R` package, subsequently.

Shiny App
¹⁵¹

¹⁵² The easiest way to inspect the data is using our shiny app provided here. The app
¹⁵³ allows to inspect all data sets or select data sets with certain specification using the filter
¹⁵⁴ box on the left. For example, if a researcher is interested in the flanker task, they may select
¹⁵⁵ all flanker data sets for closer inspection. After selection, researchers can choose between an
¹⁵⁶ overview of the included data sets, some descriptive statistics, and descriptive plots (see
¹⁵⁷ Figure XXX). If they want to further analyze the data, they can download the data via the
¹⁵⁸ "get the data tab", either directly as `csv` file or using the provided `R`-code.

Julia ¹⁵⁹ *(TODO: Add screenshot from the shiny app as figure.)*

---

[1] The newest version can be accessed via https://github.com/jstbcs/acdc-database/blob/main/acdc.db, the
Julia version at the time of submission can be found here. *(TODO: REPLACE WITH CURRENT COMMIT.)*

160   Note that all descriptive statistics in the data base are aggregated across congruency

161 conditions. This was a deliberate choice when designing the shiny app. By withholding

162 information about the effect of interest, researches can inspect and select the data based on

163 varying characteristics of the data set (including distributional properties), but remain

164 unbiased as to the most relevant outcome variable. We hope that researchers can then

165 formulate (and perhaps preregister) hypotheses about their reanalysis without much

166 hindsight bias.

167   *(TODO: Add example descriptive plots.)*   *Julia*

R-Package

169   *(SVEN: Add an overview or the R package.)*   *Julia*

```r
install.packages("acdcquery") # install package


library(acdcquery) # load package
```

Queries and Output

171   *(SVEN: Show and explain an example query with several arguments, selection of the*   *Julia*
172 *target table, etc.)*

<div align="center">

Reliability of ACDC Tasks

</div>

174   To illustrate how ACDC can be used for reanalysis, and to further assess the reliability

175 of attentional control tasks included, we provide an example data analysis. The reliability of

176 attentional control tasks auch as the Stroop task has been critically discussed in recent

177 literature (Draheim, Mashburn, Martin, & Engle, 2019; Hedge, Powell, & Sumner, 2018;

178 Rey-Mermet, Gade, & Oberauer, 2018; Rouder & Haaf, 2019; Rouder et al., 2023). Here, we

179 rely on methodological development by Rouder and colleagues @[Rouder:Haaf:2019a; Rouder

¹⁸⁰ et al. (2023)] to survey the reliability of data sets in ACDC. This analysis can be understood

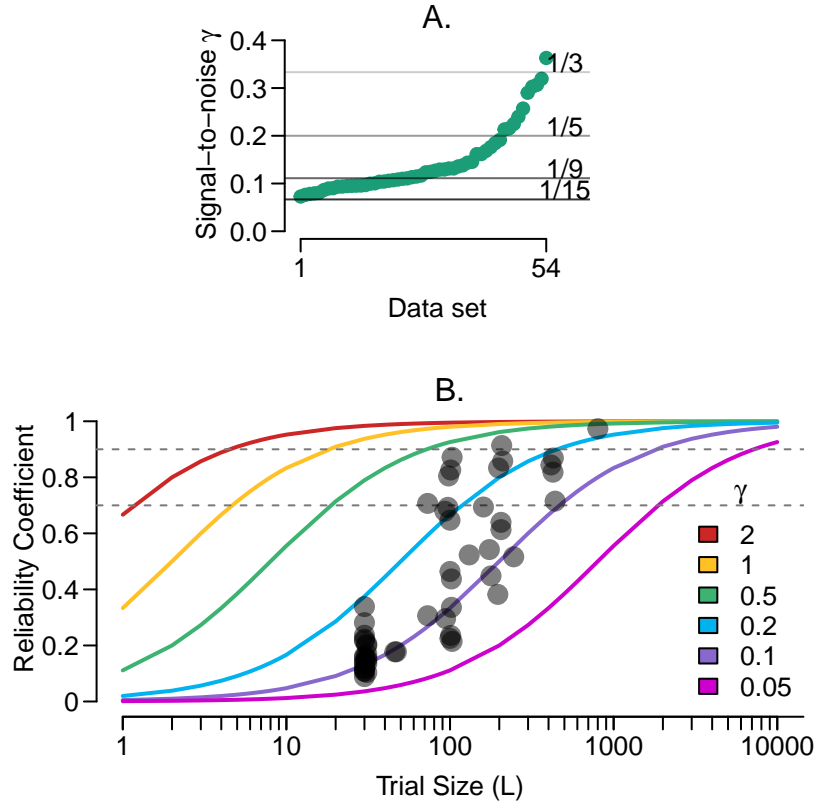¹⁸¹ as an extension of to...

Figure 3. Signal-to-noise ratio and reliability. A. Signal-to-noise ratio γ for all data sets ordered from smallest to largest. B. Reliability as a function of number of trials for all data sets in ACDC. The lines represent different signal-to-noise ratios. For a fixed γ, the relationship between the number of trials per condition and the reliability is deterministic.

¹⁸²   **Discussion**

Appendix

(INSERT TABLE WITH ALL DATASETS AND PUBLICATION REFERENCE *Julia* *HERE.)*

References

186 Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in

188 differential and developmental research: A review and commentary on the problems and

189 alternatives. *Psychological Bulletin*, *145*(5), 508.

190 Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models.

191 *Psychological Methods*, *22*(4), 779–798.

192 Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive

193 tasks do not produce reliable individual differences. *Behavioral Research Methods*.

194 MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review.

195 *Psychological Bulletin*, *109*, 163–203.

196 Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about

197 inhibition? Searching for individual and age differences in inhibition ability. *Journal of

198 Experimental Psychology: Learning, Memory, and Cognition*. Retrieved from

199 http://dx.doi.org/10.1037/xlm0000450

200 Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in

201 experimental tasks. *Psychonomic Bulletin and Review*, *26*(2), 452–467. Retrieved from

202 https://doi.org/10.3758/s13423-018-1558-y

203 Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). *Why many studies of individual differences

204 with inhibition tasks may not localize correlations. Psychonomic Bulletin and Review*.

205 Retrieved from https://doi.org/10.3758/s13423-023-02293-3

206 Sloman, S. A. (2015). Opening editorial: The changing face of cognition. *Cognition*, *135*, 1–3.

207 Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of

208 Experimental Psychology*, *18*, 643–662.

209 Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of

210 psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728.

211 Retrieved from http://wicherts.socsci.uva.nl/datasharing.pdf