

1 Attentional Control Data Collection: A Resource for Efficient Data Reuse

2 Julia M. Haaf¹, Madlen Hoffstadt¹, & Sven Lesche²

3 ¹ University of Amsterdam

4 ² University of Heidelberg

5 Version 1, October 2023

6 Author Note

7 We are indepted to Arte Bischof for her thesis work on the initial SQL data base.

8 This work was supported in part by a Veni grant from the NWO (VI.Veni.201G.019)
9 and a talent grant by Amsterdam Brain & Cognition (ABC.T09.0921) to J.M.H.

10 The authors made the following contributions. Julia M. Haaf: Conceptualization,
11 Writing - Original Draft Preparation, Writing - Review & Editing; Madlen Hoffstadt:
12 Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Sven
13 Lesche: Conceptualization, Writing - Original Draft Preparation, Writing - Review &
14 Editing.

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

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

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

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

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

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: Open Data, Attentional Control, SQL

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

researcher needs to be able to find open Stroop task data. Currently, they could either search for papers on the topic and check whether open data are provided, or search directly via OSF or other data sharing servers. However, neither of these options is very promising as the vast majority of articles in the literature does not provide raw data and data sharing servers are not equipped with sufficient search options. Second, data sets need to be accessed easily and in a general, understandable format ready for reuse. There are data sharing formats that provide this structure [REF], but they are rarely used. Additionally, data are usually shared on the level necessary for the original analysis. In case of the Stroop task, shared data might provide the Stroop effect per participant, but for this new analysis the researcher needs trial-level data. So again, there is yet another barrier for data reuse.

We think it is necessary to provide a data sharing solution that solves the current problems and enables easy and efficient data reuse. Here, we propose to gather open data sets from a specific research area in an SQL data base. This process requires little to no work in addition to current data sharing policies from the authors of original papers, some work from the lab(s) setting up the data base, and little to no work from the researchers who wish to reuse open data. We describe the process and structure we used to set up a data base of attentional control tasks called the Attentional Control Data Collection (ACDC). The data base includes XXX data sets from XXX publications from tasks like the Stroop, Simon, and flanker tasks. Subsequently, we show how the data can be explored using a Shiny app and accessed using an R-package. In an example analysis, we assess the reliability of the included tasks. This section highlights how an open data base like ACDC can aid meta-analytic efforts as well as methodological innovation.

To provide a little history of the project, the Attentional Control Data Collection was inspired by a collection of open data sets from attentional control tasks by the Perception and Cognition Lab led by J. Rouder (url). Colleagues provided the first author and Rouder with data sets for their statistical work (Haaf & Rouder, 2017; **Rouder:etal: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.

SQLite Data Base

One of the most standard ways in computer science for storing data is using SQL data bases. Structured query language (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 SQLite 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.

Data Base Structure

- SQL databases are composed of several data tables consisting of rows and columns.
 - Each row in a data table has foreign key (essentially a row ID) which uniquely identifies it. Additionally, SQL data tables may contain foreign keys which reference a unique row in another data table. In contrast to primary keys, these foreign keys allow for duplicate values within the same data table.
 - For instance, a study table may store information about all studies in a database where each row corresponds to one study. Here the foreign 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` which references the unique identifier of the respective publication in a publication table (see Figure 1).
- The structure of ACDC is adapted to the logic of publications consisting of one or

multiple studies which in turn include one or several datasets (see Figure 2).

- Each dataset stores information about a single attentional control task within a certain study. A dataset table stores information about each dataset (such as sample size) while the observation table contains the actual attentional control task data.
 - A between table contains information about between-person manipulations on a study level and a within table stores information about within-person manipulations within a dataset.
 - Note that since the congruency between stimuli is part of every attentional control task it is not considered a within-manipulation in this database but is per default included in the observation table.
 - The combination of within- and between-manipulations results in several possible conditions within a dataset, which are stored in a condition table.
- * · Add this?: Both the condition table and the observation table contain the within_id, between_id and condition_id. We deliberately chose this duplication within our database to increase the speed of accessing data through our R package and through the R shiny app.

128 **Included Data**

129 **Accessing the Data Base**

130 **Shiny App**

131 **R-Package**

132 **Queries and Output**

133 **Example Analysis**

134 **Reliability of Experimental Tasks**

135 **A Closer Look at the Stroop Task**

136 **Discussion**

References

- Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, 22(4), 779–798.
- MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163–203.
- Sloman, S. A. (2015). Opening editorial: The changing face of cognition. *Cognition*, 135, 1–3.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.
- Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, 61(7), 726–728.
- Retrieved from <http://wicherts.socsci.uva.nl/datasharing.pdf>

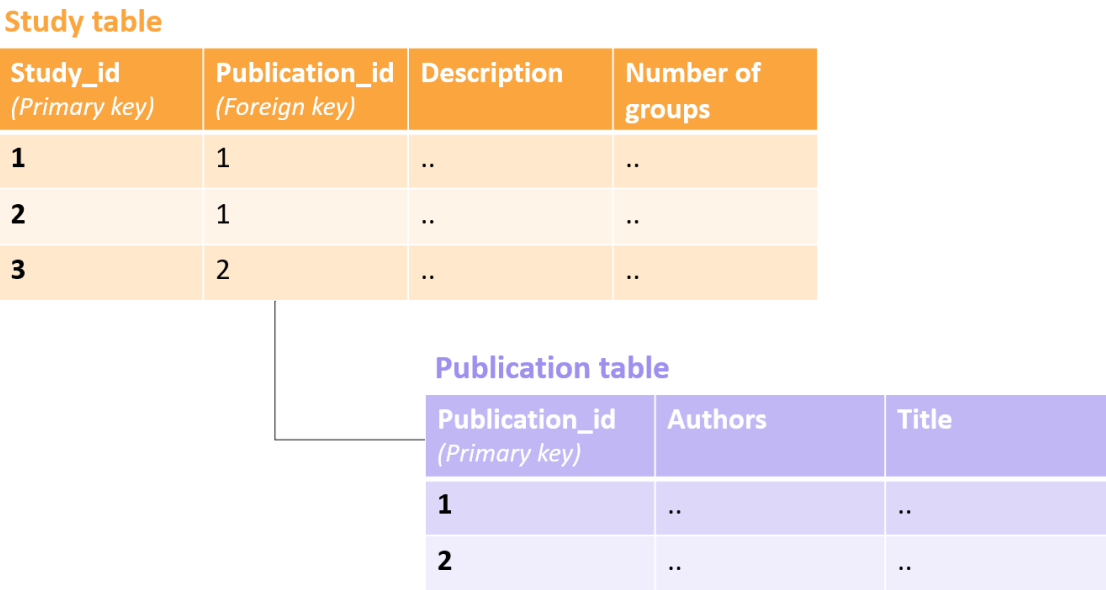


Figure 1. (#fig:figure 1)Figure 1. Illustrative example of using foreign and primary keys in a SQL database.

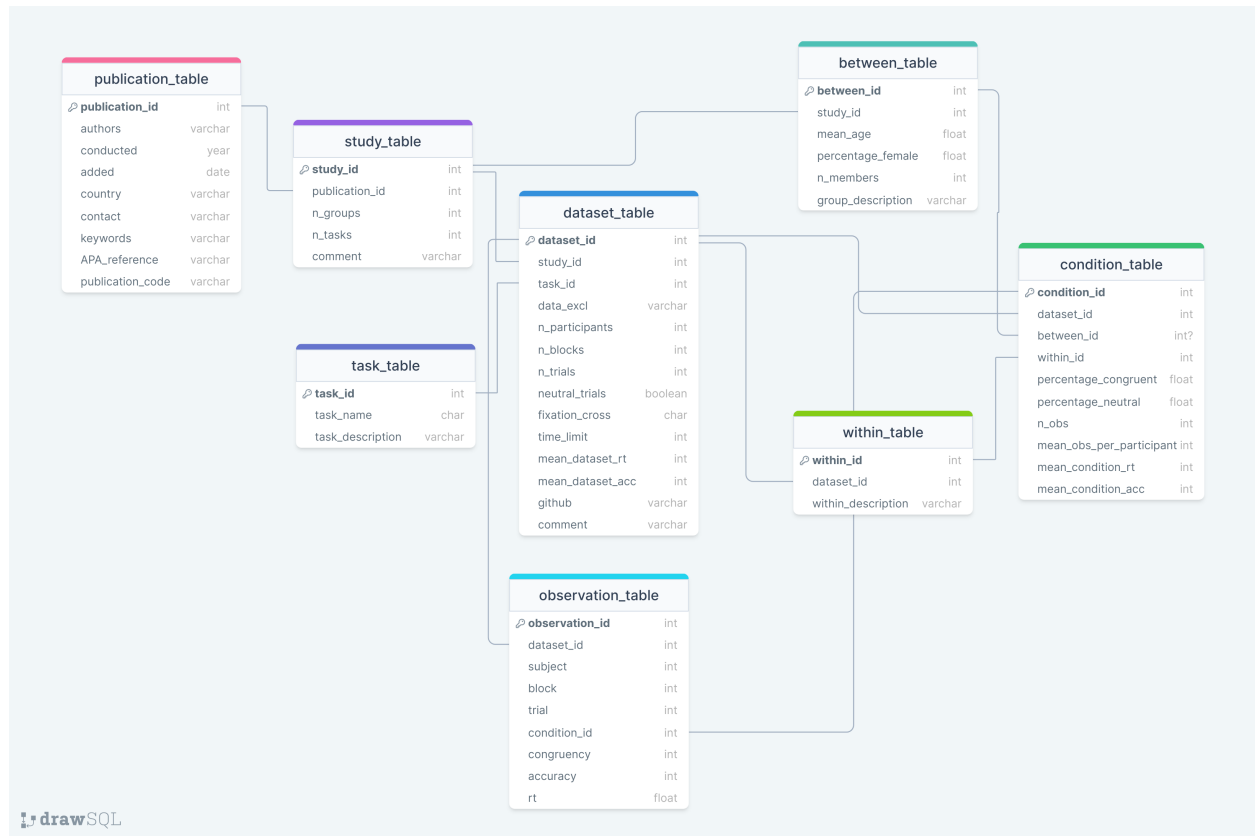


Figure 2. (#fig:figure 2)Figure 2. Structure of the ACDC database. Note. Primary keys are indicated by the key symbol. References between data tables are illustrated through lines connecting columns across data tables. This overview includes the data type of each column: integers (int), numbers with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