Challenges and Solutions for Data Analysis in an Adult Lifespan Study of over 100,000 Online Cognitive Test Completions

Annalise A. LaPlume, Ph.D.[1*]

[1] Rotman Research Institute, Baycrest Health Sciences, Toronto, Canada
* Annalise LaPlume is now at McGill University, Montreal, Canada

Declarations of interest: none

Corresponding author: Dr. Annalise LaPlume, annaliselaplume@gmail.com

Baycrest Health Sciences, 3560 Bathurst St, Toronto, ON M6A 2E1

Abstract

In this case study, I describe methodological insights from data analysis of an online adult lifespan dataset (over 100,000 completions, ages 15-100). The data were used to study cross-sectional age differences in cognitive performance. I cover the steps of data analysis for large-scale web-based data, namely data cleaning, analysis, and visualization techniques. In each step, I describe the unique challenges that face analysis of data collected online, and potential solutions to address them, by drawing on practical lessons and examples from this study.

First, I address how to identify problematic recordings such as technical issues (incomplete data, multiple completions by the same person, etc.), unreliable self-reported demographic information (age), and cognitive task outliers (accuracy, response times). I propose rigorous data cleaning as an essential first step to ensure that analytical conclusions are reliable and unbiased. Next, I demonstrate data visualization techniques that are better suited to large online datasets than more conventional techniques (e.g., density plots or locally weighted scatterplot smoothing instead of dot-plots or linear regression). Lastly, I cover the limitations of significance testing in large online datasets, and the value of complementary approaches such as data visualization, effect size estimation, and use of parsimony criteria. I also discuss more sophisticated analysis options enabled by large online datasets, such as non-linear regression, model comparison and selection, data resampling, and addition of covariates.

Learning outcomes

By the end of this case study, readers should be able to

- Identify ways to identify technical errors or false demographic reports in online data
- Apply ideal strategies for trimming online data
- Recognize techniques to visualize data appropriately in very large datasets
- Learn of the limitations of significance testing in very large datasets
- Consider alternative approaches to significance testing (e.g., data visualization, effect sizes, model parsimony comparisons)

Project Overview and Context

It is not uncommon for older adults to wonder, "Is my brain working normally? Should I see a doctor?". Perhaps they find that they lose their keys more often, or are experiencing difficulty remembering people's names. They may wonder if such memory struggles are a normal part of getting older, or if there is something they should be concerned about. When they are worried if their body is working normally, they can go to a doctor. Within minutes the doctor can use digital tools to give them readings of their body's functions (e.g., temperature, blood pressure), and tell them if they fall within a healthy range for their age and background. Wouldn't it be nice if they could get similar, quick readings on their brain? Maybe, even, without the need to see a doctor?

A team of scientists at Baycrest hospital in Canada set out to answer this question (Troyer et al., 2014). Normally, to understand how a person's brain is doing, they visit a neuropsychologist. The neuropsychologist runs a number of tests, sometimes verbal, sometimes using paper and pencil. These tests are very good at diagnosing mental concerns, but, they take time to complete, are often expensive, must be administered by a professional, are usually done in a lab or clinic, and can sometimes be uninteresting to the participant. The Baycrest science team sought to develop an online test that was quick, self-administered, easy to access, and engaging, called the Cogniciti Brain Health Assessment (Troyer et al., 2014). Designed to serve as "a thermometer for the mind" (www.cogniciti.com), people can click on a web-link to complete a short demographic questionnaire, followed by online versions of psychological tests measuring memory and attention. At the end, they are given a percentile score, which tells them how to they fared compared to others of the same age and educational background, and whether they should see a doctor for follow-up.

Online assessments are growing in popularity, and they have made psychological tests far more accessible and available to the public. The Cogniciti assessment is unique for several reasons. First, it was specifically developed for older adults. As such, factors such as age-appropriate stimuli, and familiarity with computers were taken into account when designing and extensively piloting the web-based platform (Troyer et al., 2014). Second, it was psychometrically validated, which is rare for online tests, and confirmed that the test consistently obtained the same results when repeated (reliability), and that it measures what it attempts to measure (validity; Paterson et al., 2021; Troyer et al., 2014). Lastly, it was developed as a screening tool for memory concerns, and reliably distinguishes people who are 'worried well' from those who should be seeing a health care professional. To do this, the test developers established a range of normal performance, using tests of cognitive abilities that are known to recruit brain regions influenced by aging and age-associated cognitive disorders.

The assessment has been available online since it was validated in 2014. In the following six years (2014-2019), it was completed over 100,000 times. These completions were from people who heard about the test online, in media, or from a friend. While the goal was to offer a free and reliable screening tool, test-takers also consented to have their data anonymously stored. In the current study, I, along with a team of researchers, leveraged this unparalleled dataset (LaPlume et al., 2021). Our goal was to map cross-sectional differences in how and when cognitive performance differs from one age to another.

Section summary

- A unique online measure was used, the Cogniciti Brain Health Assessment, that was designed and validated as a screening tool for older adults with memory concerns.
- Data for this case study are from data collected from online test-takers over six years.

Research Design

Using the large online dataset and statistical modelling, our research team was able to study cognitive aging in a more fine-grained way than past lab investigations. In this case study, I describe methodological insights from analysis of the web-based dataset collected with the Cogniciti Brain Health Assessment (115,973 completions, ages 14 to over 100; LaPlume et al., 2021). Details of test development and validation have already been covered in detail (See Troyer et al., 2014).

Online technologies have revolutionized cognitive aging research, as they allow collection of much larger sample sizes from different ages, which then lends itself to more sophisticated statistical analyses of aging effects (Hartshorne & Germine, 2015). However, using online data, which typically involves very large datasets, also yields unique analytic challenges for the researcher. I will go over these challenges, and methodological decisions made in response to each challenge.

Section summary

- The goal of the research study was to study cross-sectional differences in cognitive performance across age.
- This case study focuses on the methodological challenges, and possible solutions, when analyzing large-scale online cognitive lifespan data.

Method in Action and Practical Lessons Learned

In this case study, I focus on the data analysis aspect of the research, along with the related steps of data cleaning and visualization. I go over the practicalities of each step of data analysis, and then draw lessons learned from each step, highlighting insights that other online researchers can apply to their studies. I focus on general insights for analysis of large-scale online datasets, with additional insights on the analysis of cognitive data from people of different ages.

Data Cleaning

Before data can be analyzed, it must be carefully processed and examined for quality. Data cleaning is time-consuming, but is a good investment. The colloquial phrase "Garbage In – Garbage Out" recognizes that entering poor quality data into an analysis results in unreliable data output (Kilkenny & Robinson, 2018; Rose & Fischer, 2011). Data cleaning is especially important for online data collection, which is less controlled than data collected in a lab or clinic. Unique challenges face online data collected not for a particular study, on social media platforms (Kim et al., 2016), or via paid research participant sites such as Amazon's M-Turk (Buchanan & Scofield, 2018; Crump et al., 2015).

Data cleaning often is the longest and most complex part of analyzing online data. As an example, in the current study, data cleaning took over twice as long as data analysis. Data cleaning was on the scale of several months in the current study, and can even take several years for complex datasets (e.g., Farhan et al., 2016; Sunderland et al., 2019).

Increasing work has shown that online testing can have comparable reliability to lab testing (Chetverikov & Upravitelev, 2016; Crump et al., 2013; de Leeuw & Motz, 2016; Hilbig, 2016; Reimers & Stewart, 2007, 2015; Slote & Strand, 2015), but this is not always the case (Feenstra et al., Germine et al., 2019; Miller & Barr, 2017; Parsons et al., 2018). Data cleaning is crucial to ensure that online data is reliable, given the loss in control during web-based testing

(Chetverikov & Upravitelev, 2016). The good news is that the large sample sizes that are easily obtained online allow for more rigorous data cleaning, as it is less expensive for the researcher to drop data from online participants than to drop data from lab participants tested in-person.

Section summary

- Data cleaning is fundamentally important for all research as it determines the reliability of results from data analysis.
- Data cleaning is especially important in online studies, due to the less controlled testing environment.
- The large sample sizes from online data collection lend themselves to rigorous data cleaning.

Technical Errors

Our first step in data cleaning was to look for technical errors. I removed completions if participants refreshed the page during the task (*n* = 113), or had technical issues with data recording that lead to incomplete test records (*n* = 15,541). A single odd case I observed was completion of the test by what appeared to be artificial intelligence, in which the test was attempted 12 times in one second. By removing incomplete data, I restricted ‘casual’ participants, as described by Dr. Stian Reimers, individuals who wanted to see what the assessment was like but did not mean to provide proper results for analysis (Reimers, 2007).

Unless you are collecting longitudinal data, it may be necessary to remove additional completions by the same person, to avoid learning or practice effects and ensure independence of observations. As such, a given participant’s data from different occasions must be linked in a way that does not compromise confidentiality. In the current dataset, participants were assigned a unique alphanumerical ID based on the email address they used to sign up for the test.

Subsequent completions by the same ID were removed (*n* = 20,687), with the assumption that one email address represents one person. In other cases, researchers have recorded a computer's I.P. address to create a participant ID, and while this avoids having to collect personal information such as an email address, it means that different people using the same computer are recorded as the same person. In addition, I.P. addresses can be used to identify people, and so it is better not to record them.

One of the biggest differences between working with large online datasets compared to smaller datasets is that a researcher cannot simply scroll through the collected data to look for inconsistencies. Instead, I suggest running spot-checks on the data, to identify if randomly selected entries appear to be consistent. I also suggest programming scripts to examine the data, calculate frequencies and ranges for observed variables, and compare the observed values to those expected. I used the tidyverse suite (Wickham et al., 2019) in the R statistical environment (R Core Team, 2020), as it offers a collection of packages and functions for manipulating and dealing with large datasets (e.g., the *glimpse* or *head* functions to view data samples). Writing scripts for data cleaning also enables you to create programming pipelines that provide a record of the cleaning steps taken, which can be updated for new datasets.

Section summary

- Technical errors of refreshing the page, incomplete completions, or multiple completions, should be considered for removal.
- Creating scripts is a good way to track and reproduce cleaning steps.

Demographic Inaccuracies

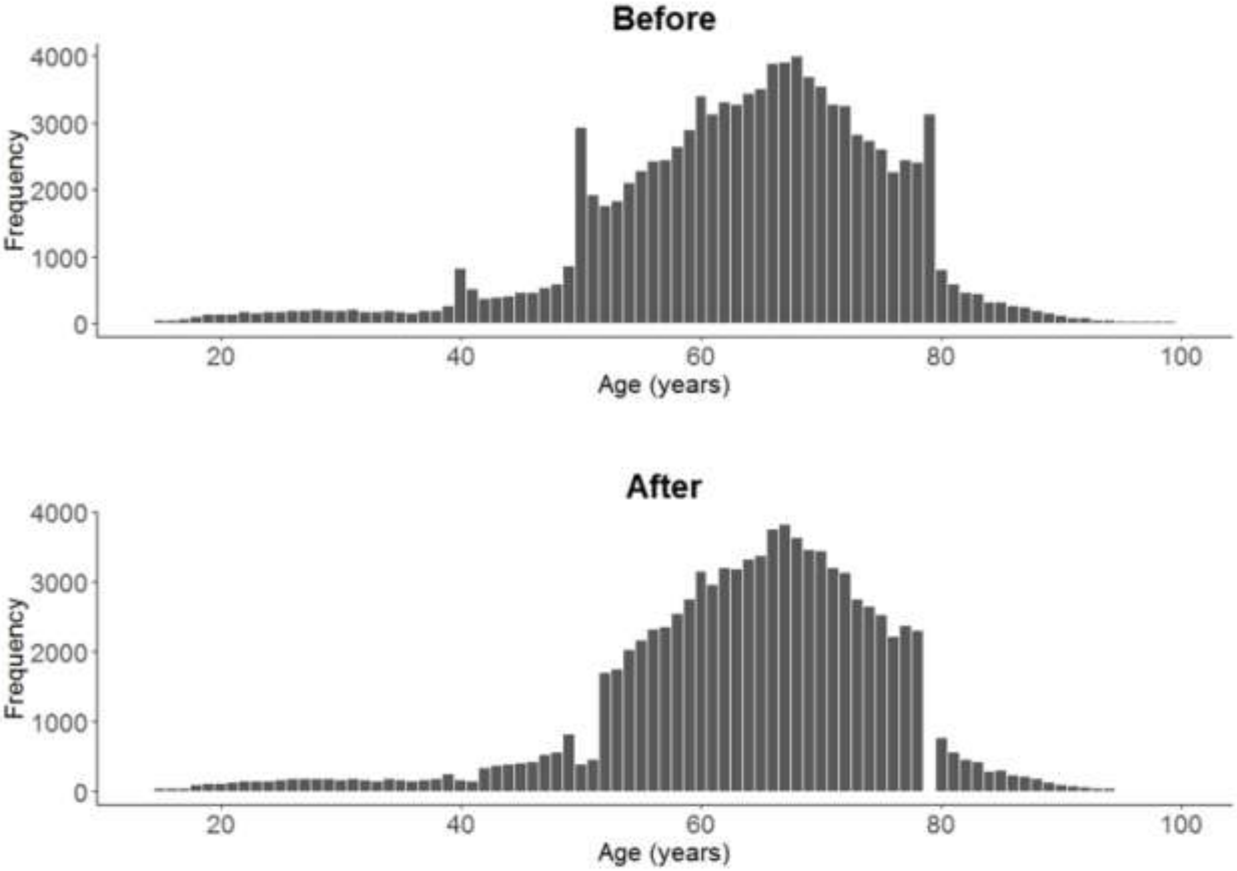
A general lesson on data cleaning is that attention must be paid to how particular steps of your data collection methods can compromise the data collected, and ultimately the results obtained. Different methods of data collection can capture vastly different data on the same research questions. Thus, in each data cleaning step, a fundamental general lesson is to attend to specific errors or inaccuracies that could arise from your particular data collection methods.

It is especially critical to attend to data inaccuracies that compromise how the available data answers your particular research questions. For example, in our case, participants would have been motivated to lie about their age to receive a test score. As age was the primary variable of interest in the current study, any inaccurate ages reported by participants would have led to biased or misleading conclusions. Age was self-reported, and the anonymous nature of the test meant that I did not have a means to independently verify the reported age. However, I did find several creative ways to identify seemingly inaccurate ages.

First, I removed ages that were extremely unlikely to be accurate, namely individuals above 100 years old. I also removed ages below 18 and above 90 (*n* = 564), because there were so few (below 50 people per age), and because people in those ages did not perform as expected for their age when compared to others of nearby ages, indicating that those individuals may not have been truthful about their age.

Next, I considered motivations for people to provide incorrect ages. The online assessment that I used was designed for older adults. In the initial years of data collection, the test was validated for ages 50 to 79 (2014-2016), ages 40 to 79 in the next few years, (2017-mid 2019), and for all ages halfway through the final year (mid 2019). Although people of any age could take the test, a score indicating how well they performed was only provided to people at validated ages. After participants had entered in their age online, they were shown a warning stating they could complete the test but would not get a score unless their age fell within a given range. For this reason, test takers may have been motivated to provide a (false) age within the range that would give them feedback. Indeed, when viewing the data, I observed a disproportionate spike in frequencies per age, with twice as many people who reported their ages exactly at or near the cut-offs of the validated age range (“40”, “41”, “50”, “51” or “79”) compared to nearby ages (Figure 1).

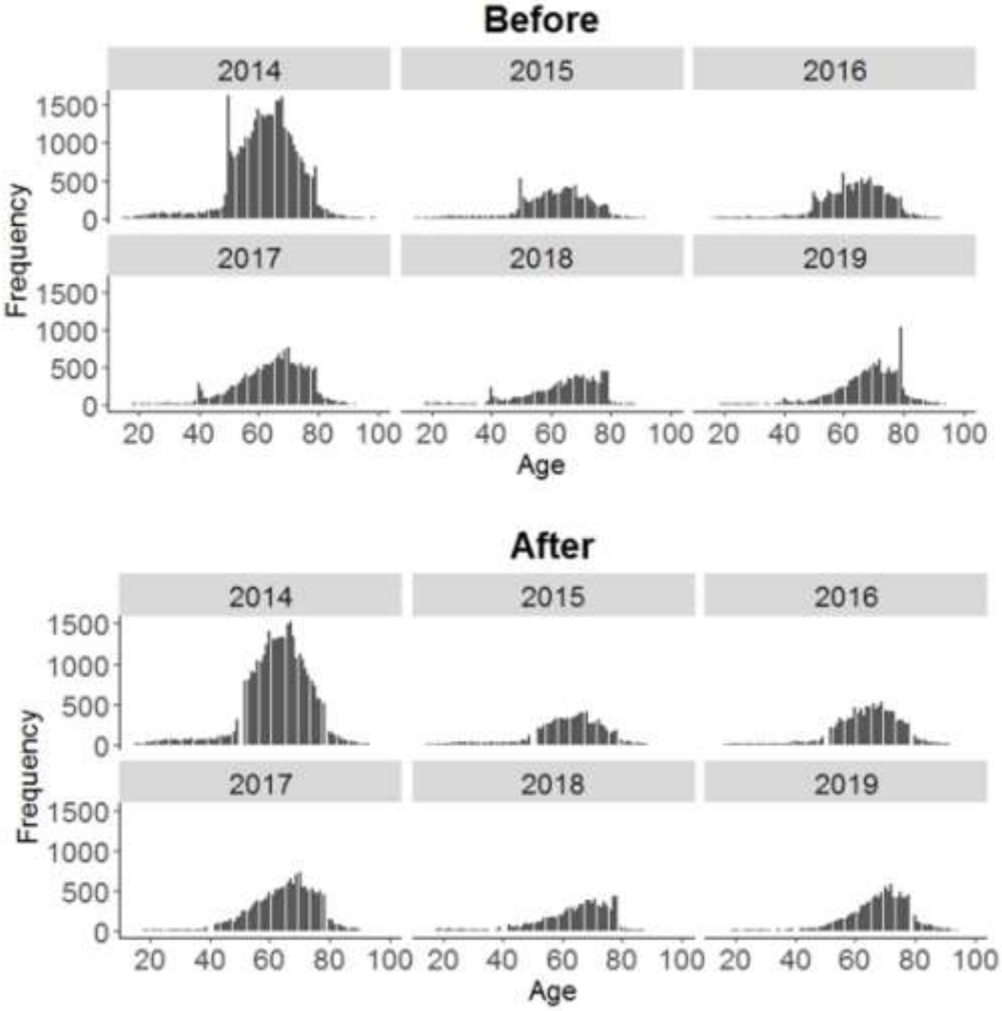
Fig 1. *Histograms showing frequency of test completions per age, before and after data cleaning*



When I separated the age frequencies by year, I noticed that the spike in certain ages coincided with the cutoffs for each year, and that spikes disappeared after all ages were validated

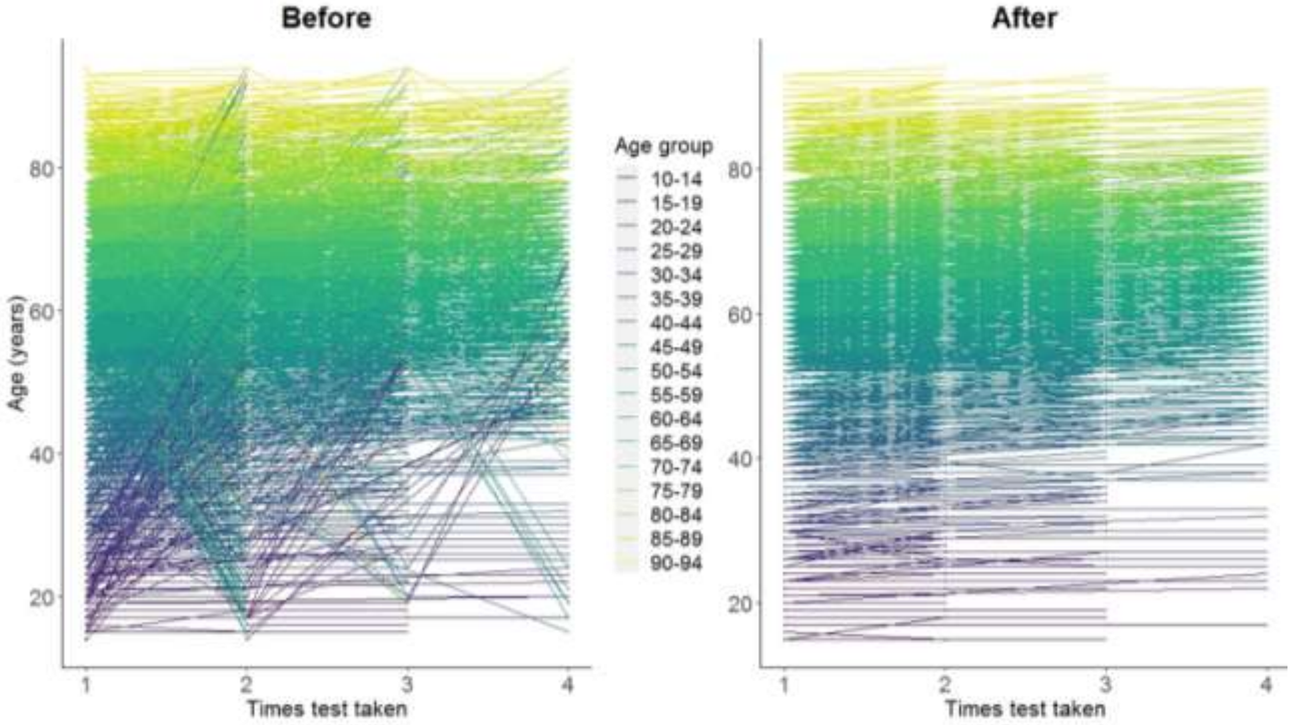
(Figure 2). The performance of people at the cut-off ages also deviated considerably from those of nearby ages. As I could not identify which ages were truly or falsely reported, I removed all people at the cut-offs per year (*n* = 8,093), as information from those ages was more likely to be unreliable than helpful.

Fig 2. *Histograms showing frequency of test completions per age per year of data collection, before and after data cleaning.*



A last way of determining false ages was to look for a sudden change in a person's age from one test completion to another. Spot-checks of people with several test completions revealed that some of them genuinely may have accidentally misreported their age. For example, one gentleman's score varied from '61' to '60' to '62' over three years. Mistaking one's age by a year or two is normal. However, other people had jumps in their age of 20 years, and it is extremely unlikely that a person would think they were aged '20' one day and '40' the next (Figure 3). I therefore excluded data if a participant's age changed by over five years (*n* = 4,078) in a repeated assessment of five years or less. This cut-off balanced allowed for real mistakes, while excluding deliberate misinformation.

Fig 3. *Line plots showing a person's self- reported age over the first four test completions, before and after data cleaning. As completions were within six years, any lines that are not flat or slightly sloping upwards (i.e., any sudden jumps in the lines) indicate that a person may be falsely reporting their age.*



Note. Each line indicates one person. Lines are colour-coded by age group.

Section summary

- Participants may have been motivated to lie about their age in order to receive a test score.
- Although participants' ages could not be independently verified, creative ways were employed to see if self-reported ages were consistent with other information available

Cognitive data trimming

The cognitive tasks that I used measured performance either via speed (a person's response time per trial of a task, or their total completion time for a task) or via accuracy (the percentage of errors a person made across trials of a task, or the number of clicks it took them to complete a task). With speeded data, I removed data from participants who were exceptionally fast (as this could indicate guesses or randomly clicking through the task) and exceptionally slow (as this could indicate a lapse in attention or an interruption). With accuracy data, most participants fell within an expected range of performance (especially on percentage error rates, in which the range was limited), but data were examined for any ceiling or floor effects.

To obtain reliable results, I trimmed the data twice. First, I ran within-person trimming (per person per trial of a single task), to look for outlying responses on individual trials; that is, exceptionally fast or slow responses. Next, I ran between-person trimming (per task), to look for outlying people on individual tasks; that is, exceptionally fast or slow people. Importantly, as I expected age to affect a person's speed and accuracy, the trimming was done separately for each age, to identify outliers per age.

One way of trimming cognitive data is by using an absolute response time cut-off, usually determined by the researchers’ intuition. For example, I removed people who responded faster than 250 ms, as this is a typical cut-off used in cognitive studies trying to determine the minimum amount of time it takes to actually make a decision. Another trimming method is to remove response times that fall outside a set range. A widely used metric for a set cut-off is three standard deviations above or below the mean. However, the standard deviation is influenced by the sample size (Van Selst & Jolicoeur, 1994). A wider cut-off is recommended for smaller sample sizes, and a smaller cut-off for larger sample sizes. This is a particular issue for online datasets which have atypical sample sizes for two measures that are used in trimming: the number of participants and the number of trials. Online studies have a larger number of participants than normal, as more data can be easily collected than lab studies. By contrast, online studies have a fewer number of trials than normal, as online measures typically have fewer trials (i.e., are shorter) than lab tasks, to encourage more people to participate.

Therefore, it is better to pick a cut-off based on the sample size (i.e., number of trials or people) rather than an arbitrary value such as three standard deviations. Moreover, it is ideal to use an adaptive cut-off criterion to adjust for the change in the standard deviation as outliers are removed. I used an iterative trimming script with a moving cut-off criterion, developed by Dr. James Grange and implemented in the R package trimr (Grange, 2015). In each cycle, the slowest response is temporarily removed, and the mean and standard deviation are then calculated and used to determine the cut-off. The response is then returned to the sample and removed if it falls outside the cut-off. The process is repeated until no outliers remain.

Section summary

- Trim cognitive data both within-people (to remove problematic trials) and between-people (to remove outlying information).
- Impliment iterative trimming, using a procedure that constantly recalculates the standard deviation cut-off, to produce results unaffected by sample size.

Cognitive data reliability

One way to examine the reliability of cognitive data is to compare the obtained data to similar past studies. Our findings were similar to findings from lifespan lab datasets that measured the same cognitive abilities as in our study, namely working memory (Borella et al., 2008; Chiappe et al., 2000; Myerson et al., 2003), processing speed (Salthouse et al., 2000), interference control (Rey-Mermet & Gade, 2018), associative recognition memory (Bender et al., 2008), and set shifting (Periáñez et al., 2007; Salthouse et al, 2000; Tombaugh, 2004). Similarly, our results matched past studies indicating both a general factor of age-related variance shared across tasks, as well as specific variance on each task (Salthouse, 2017). Finally, our results matched those of increased inter-individual and intra-individual variability with age (Dykiert et al., 2012; Hultsch et al., 2002).

Moreover, the means and standard deviations that I obtained for each age half-decade replicated those from the original study using the same online cognitive assessment (Troyer et al., 2014). Together, these similarities increase our confidence in the cleaned data. They indicate, as found for previous online data collection, that unreliable data appear add random noise to the data, but do not create systematic error (Reimers, 2007).

Section summary

- Reliability of data is assessed by comparing to similar data collected on lifespan cognitive performance in the lab.
- Our collected data replicated earlier findings from the same online cognitive assessment.

Data Visualization

Data visualization serves two purposes in research studies. First, researchers should plot their data to *explore* it, by using rough figures to get an early look at the data. Second, researchers should plot their data to *present* their final analysis results, by using polished figures to convey study results to a specific audience, such as in a research paper. The first purpose takes place before data analysis, the second after data analysis. The audience of the second purpose is the reader of the final research paper, while the audience for the first purpose is the researcher themselves.

In online studies, the use of data for exploration is important because as stated earlier, it is not possible to identify trends or problematic observations by merely scrolling through the data, but these are often made clear through early data visualization. Using data for presentation is equally important because carefully selected data visualization procedures can often succinctly present available information without making any assumptions about the underlying distribution, unlike in data analysis (typical Inferential Statistics or significance testing). Further, typical data analysis aims to makes predictions (inferences) about the underlying population from which the research sample is drawn, but such inferential testing is less important for online research, as the large sample sizes can appear to be similar to a population.

Section summary

- Data visualization is useful for the purpose of data exploration before analysis, to help the researcher explore datasets that are too large to scroll through.
- Data visualization is also useful for the purpose of data presentation after analysis, and can be a valuable complement or replacement to significance testing.

Hexbin or 2d histogram plots of raw data

For data exploration, I visualized the raw data with a scatterplot of overall performance by age (Figure 4). The scatterplot allowed me to see the frequency of responses per age, the range of performance, and any extreme observations. When looking at the scatterplot, I also gained a new insight—that cognitive performance declined as age increased, but that cognitive performance also became more variable from one person to another. Some 60-year olds seemed to have similar performance to 20-year olds, while others were much worse. Our team then further examined this insight in our analyses.

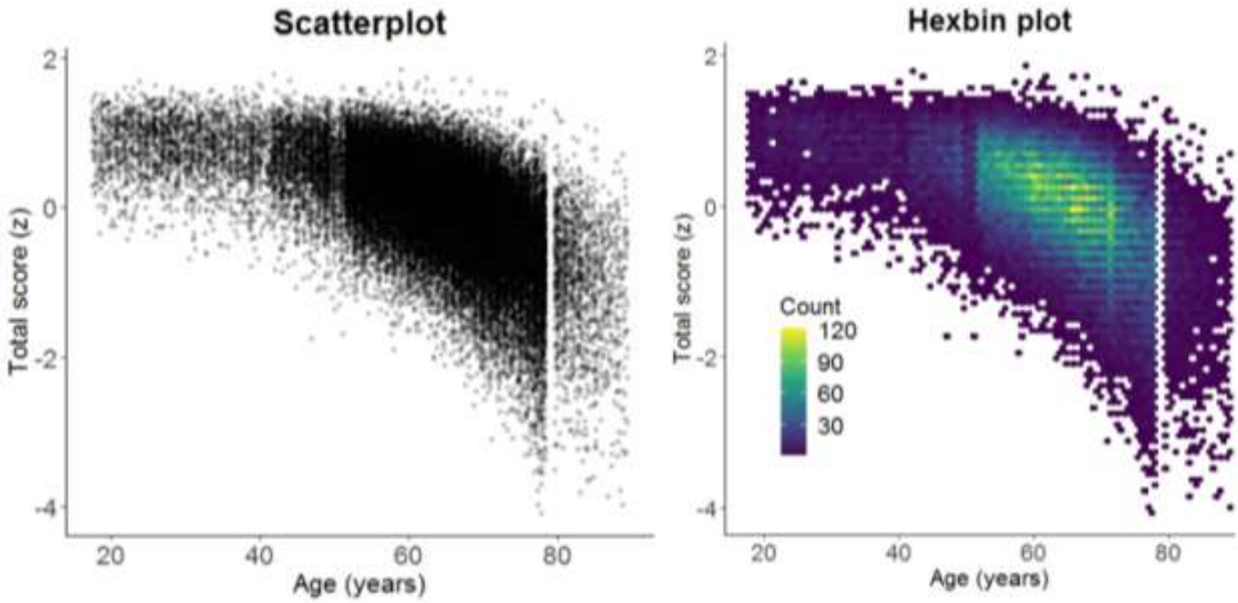


Fig 4. *Scatterplot and Hexbin Plot of Total Cognitive Performance (z-scores) per Age.*

Note. Colour-coded hexagons in the hexbin plot indicate the frequency of responses for each age.

However, the sheer volume of data leads to overlapping points in the scatterplots. To alleviate the obfuscation of overlapping points, , I used hexbin charts in R. The hexbin charts added a third dimension to capture the frequency of information available for each age (Figure 4).

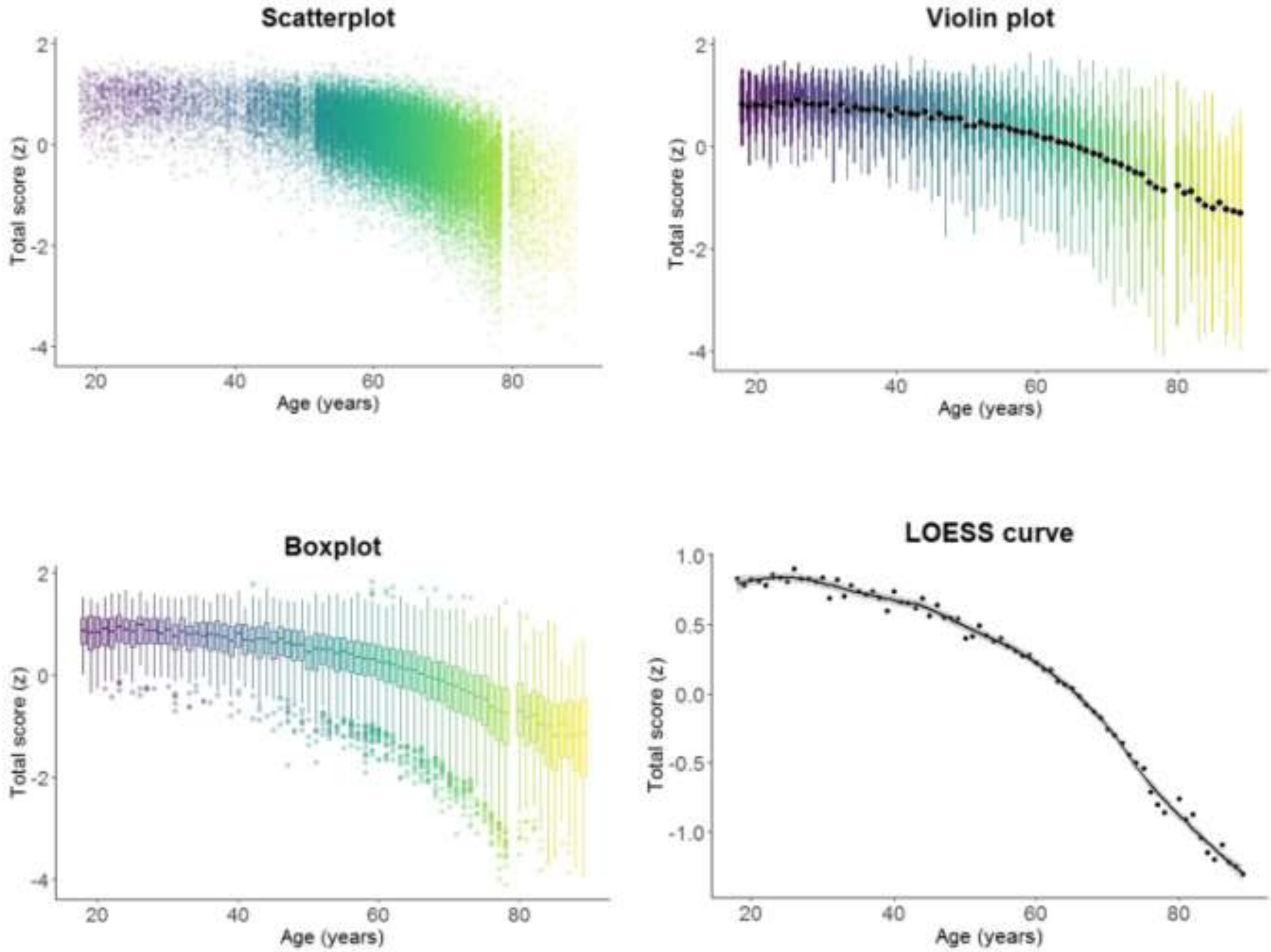
Section summary

- Exploring the data in advance confirmed our original hypothesis (that cognition declined as age went up), and also led us to a new insight (that cognitive performance became more spread out as age went up), which we then examined further in data analysis.
- Consider visualization techniques that indicate the frequency of raw data rather than individual data points, due to the volume of the data (e.g., hexbin charts or 2d histograms instead of scatterplots).

Scatterplot smoothing to summarize trends in data

For data exploration, I colour-coded scatterplots per age (Figure 5). As scatterplots only coarsely examine age trends, I also used boxplots and violin plots to more closely examine age trends via the average performance (medians or means) and spread (quartiles and range) per age. It is much easier to understand trends from these distribution plots than it is from looking at a table with the same values. For data presentation to draw the readers eye to general trends, I then employed a technique called locally estimated scatterplot smoothing (LOESS) or locally weighted scatterplot smoothing (LOWESS; Cleveland, & Devlin, 1988).

Fig 5. *Graph Types used to examine the Trend of Total Cognitive Performance (z-scores) per Age.*



Notes. LOESS=Locally Estimated Scatterplot Smoothing
Ages are coded into different colours, to help differentiate each year of age. In the boxplots, the centre line indicates median cognitive performance per age, and the dots indicate outliers. In the violin plots, the black dots indicate mean cognitive performance per age. In the LOESS curves, the black dots indicate mean cognitive performance per age, the line shows a smoothed curve with span of 0.3, and the grey shading around the curve indicates a 95% confidence interval envelope.

LOESS curves fit models to localized subsets in the data, in this case by using a moving window to examine local trends in the mean for each age. LOESS curves are a non-parametric technique, which means that they allow the researcher to present overall patterns in the data without making any assumptions about the underlying function. They allow us to clearly see trends of age differences in cognitive performance, thus targeting our main research question, without needing to determine *a priori* whether to expect a linear or quadratic function that underlie these trends.

LOESS curves would be adequate to answer our research question, as they summarize the age effects in the data. Nevertheless, I also fit parametric models to quantify age effects. The LOESS curves also helped towards this next step, as they allowed us to determine that a non-linear function would be ideal for further statistical analysis. After fitting parametric models, I could also then compare them back to the LOESS curves, to see how well parametric models fit the data.

Section summary

- Experiment with a series of different graph types (See Figure 5).
- Locally weighted scatterplot smoothing offers a good way of examining trends in large datasets without making distributional assumptions.

Data Analysis

The large sample sizes that arise from online testing mean that the application of standard significance tests is limited. As stated by Fan and colleagues, ‘‘large sample size leads to highly precise estimates and thus very low p-values for almost every effect investigated, regardless of its size, theoretical significance, or importance” (Fan et al., 2020). Thus, I chose to interpret visualizations of the data, and to use effect sizes and parsimony criteria to compare fitted models alongside statistical significance testing. At the minimum, an easy approach is for researchers to interpret effect sizes and data visualization rather than p-values. More advanced researchers (at the postgraduate level) could also apply the techniques described in the following two sections.

Model Fitting, Comparison, and Selection (Advanced)

As there is so much information available with large datasets, it is ideal to fit and compare different types of models, to find the one that offers the best fit to the data. More sophisticated models can be fit to very large datasets. For our research question, I decided to fit more than one term in the model equation (i.e., more than one parameter) to account for our variable of interest (i.e., age). First, I looked at a basic linear model predicting that age influenced cognitive performance, a simplified equation of which is presented below.

$$\text{Cognitive task score} \sim \text{Age}$$

The LOESS curves indicated that a simple linear relationship may not be adequate, so I considered some non-linear alternative models. I tried quadratic models with linear and polynomial terms for age,

$$\text{Cognitive task score} \sim \text{Age} + \text{Age}^2,$$

and one with a cubic term for age

$$\text{Cognitive task score} \sim \text{Age}+ \text{Age}^2 + \text{Age}^3{}_.$$

The best model so far was the quadratic model. I next tried a segmented regression model (also called changepoint or piecewise regression), which accounts for a ‘bend’ (a segment or changepoint) in a linear relationship,

$$\text{Cognitive task score} \sim \text{Age} + \text{Age changepoint}.$$

A downside of having so much information available is that a researcher may select a model that perfectly captures all the information in the current dataset, but, because this model is perfectly tailored to the available data, it does not then generalize to other datasets. In statistics this is called ‘overfitting’, and happens when a model fits a set of obtained data too closely or exactly, and cannot fit or predict future data reliably. I thus recommend careful model comparison and selection when choosing a final model.

Simpler models are less prone to overfitting than complex models, as they involve fewer explanatory variables (parameters). At the same time, more complex models can be useful in providing a more detailed picture that capitalizes on the information present. However, an important question is whether the additional detail is necessary. The goal when comparing models is to achieve parsimony, that is, to select the simplest model that explains the data, and thus model only necessary complexity.

To compare between models, I used hierarchical regression to see if each more complicated model offered a statistically significantly better fit than a simpler model, using a chi-square difference test to compare between models. For example, I compared if the model with age plus one changepoint was significantly better than the simpler model with age only. However, the utility of significance testing alone was limited due to the size of our dataset, and hierarchical regression often revealed that complex models offered a statistically better fit to their simpler counterparts.

To complement significance testing, I then examined whether more complex models produced a change in effect size. Using regression models, I measured effect size as the change in variance explained (r^2), and looked for models that explained at least 1% more variance (indicating a small effect size according to published guidelines; Cohen, 1988). By adding effect sizes, I could assess if a more complex model added practical value.

In addition, I assessed model parsimony with the Akaike Information criterion (AIC; Akaike, 1957) and the Bayesian information criterion (BIC; Schwarz, 1978). The AIC and BIC have different properties and assumptions, hence it is ideal to examine both. The BIC is more conservative, and corrects for sample size, making it ideal for very large samples. It is also consistent for model selection with large sample sizes, because as the sample size increases, the likelihood of it selecting the true model increases. However, the BIC aims to select a true model, thus it assumes that a true model with measurable values exists, and that the true model is in the set of candidate models. Hence the AIC is better if it is unclear whether there is a true model or if the true model is too difficult to specify via an equation.

Often these criteria diverged (p-values, r^2, AIC, BIC). In those cases, I used the most conservative criterion. For the criteria I used, p-values were the least liberal, followed by the AIC, and finally the r^2 and BIC.

Section summary

- It is ideal to fit different models to a large dataset, and compare them to select the best-fitting model.
- When selecting a final model, consider effect sizes and parsimony criteria alongside significance testing.

Covariates (Advanced)

Another way of modelling the data in more detail is to include covariates, variables that could influence the measure of interest in addition to the target variable. In our case, I added covariates for variables that I expected would influence cognitive performance in addition to our predicted measure of age, namely a person's sex, level of education, the time-of-day in which they took the test, and the year in which they took the test, as follows:

Cognitive task score ~ Age + Sex + Level of education + Time-of-day + Year

I also predicted interactions between age and each of these variables, as follows:

Cognitive task score ~ Age + Sex + Level of education + Time-of-day + Year +

Age*Sex + Age*Level of education + Age* Time-of-day + Age*Year

Note that it is not always appropriate to include covariates, even when covariates may be expected to influence results, unless there is an ideal sample size for each additional covariate. Having a small sample size for a covariate will lower the ability to detect an effect that does exist (i.e., power), as power is based on the group with the smallest sample size. Also, covariates are mistakenly considered to 'control for' group differences on the independent variable (e.g., adding level of education 'controls' for possible differences in education levels for people of different ages), but covariates should actually be expected to be equally distributed over the independent variable, and should instead be used if they account for variance in the dependent variable (Miller & Cohen, 2021).

I independently tested whether each covariate should be included, using purposeful variable selection methods (Hosmer et al., 2011). This technique balances between variables that account for variance in the outcome, versus variables that have minimal predictive power and produce an unnecessarily complex model. Including all confounding variables (even non-significant ones) can increase standard errors and produce instability in estimated values. I used regression models to separately examined the unadjusted relationship between each covariate and the outcome (e.g., Cognitive task score ~ Sex), and then the adjustment on age by each covariate (e.g., Cognitive task score ~ Age + Sex). The end result was a multivariate linear model that regressed age on the task score, and included covariates and covariate interactions that were shown to have significant effects on the outcome.

I found that adding covariates to a model with the variable of interest actually helped avoid over-fitting compared to models fitted to the variable of interest alone. As such, the best-fitting model with age alone tended to be unnecessarily complex, with several different terms for age, such as either a third degree polynomial or a two-changepoint model. Side-by-side comparison with LOESS curves indicated that these models were over-fitted, and were modelling variance due to noise in the data. In comparison, once covariates were included, the best-fitting model tended to be less complex, such as a second degree polynomial or one-changepoint model, and appeared to model overall age trends more appropriately. Therefore, the covariates accounted for some of the variance that was being misattributed to effects of age.

Section summary

- Including covariates can help explain some of the variance in the data.
- Effects of individual covariates can be assessed before including them in the model.

Assess data quality/representativeness

In addition to screening for low-quality data, it is recommended that researchers try to quantify the effects of data quality, such as by analyzing results with and without flagged items (Tweedie et al., 1994). This is especially useful in cases where a researcher may not wish to remove all of the low-quality data. In our case, one issue was that although there was a very large sample of different ages, the sample was not equally distributed across ages (Figures 1 and 4). The mean age was 63, with a standard deviation of 12, indicating that the majority of participants were in their fifties, sixties, and seventies.

While there still were at least one thousand people per age decade, I wanted to test if this unequal age distribution biased our results due to the over-representation of some ages and the under-representation of others. To do this, I created a subset of data that included equal samples per age. I did this by examining the smallest sample available for each single year of age (*n*=50), and then randomly sampling the same sample size for the other years of age. This created a subset of the original data, in which all ages were equally represented. I did this a number of times, to create several randomly sampled subsets. I then compared the results from these subsets to the complete dataset. Although findings were similar for both, the results for the best-fitting models from the equal samples subset were a bit more conservative. I thus used the best-fitting models from the equal sample dataset, to control for any bias from uneven sample sizes.

I also tried another random sampling technique. Instead of using the smallest available subset per age (*n*=50), I used an ideal subset per age (*n*=100), and then resampled data for ages in which there were fewer people than this number. This produced similar results to the first sampling option. Researchers could also try cross-validation resampling techniques of splitting the data into ‘test’ and ‘training’ datasets, and seeing how well models fit on the training dataset predict the test dataset. Another option to deal with uneven sample sizes is to weight the data by the sample size, in this case, using inverse weights of the sample per age.

Section summary

- Random samples can be obtained from the data, and the results compared for each random sample.
- I used a random sampling technique to create a representative subsample to control for any bias from uneven sampling.

Conclusion

Web-based testing offers a challenging but opportunistic frontier for research. It is especially important when in-person research is no longer possible, which has lead to its rise during the world-wide COVID-19 pandemic that began in 2019. Online assessments have the dual benefit of making research tools freely available and easily accessible to the public, while allowing researchers to collect large sample sizes. Thus they facilitate ‘citizen science’ (Germine et al., 2019), by lowering barriers for the public to participate in scientific research.

Increasing work has examined the validity and reliability of developing or using online assessments (Chetverikov & Upravitelev, 2016; Crump et al., 2013; de Leeuw & Motz, 2016; Hilbig, 2016; Reimers & Stewart, 2007, 2015; Slote & Strand, 2015). In this case study, I outline the unique challenges of analyzing online data, as well as potential solutions. Extensive data cleaning is a necessary first-step before data analysis, and data visualization is a powerful tool to accompany statistical testing. I recommend use of measures such as effect sizes and parsimony criteria, in addition to conventional significance testing. Online data collection offers many benefits for studying human behaviour, but it is important to properly consider data analysis steps, and to look creatively beyond conventional data analysis methods.

Discussion Questions

1. Imagine you are designing an online test. What are strategies you could use to increase the accuracy of the data that will be collected?
2. Why is data cleaning important in web-based studies? What are some accidental or intentional errors that can arise when people complete online assessments?
3. How can you check the reliability and validity of data collected from online sources?
4. What the two purposes of data visualization in online research? Can you think of any other visualization techniques that would be suited to large online datasets?
5. Why is overfitting a problem in online data analysis, and how can researchers prevent it?

Multiple Choice Questions

1. What is the principle of parsmony?
    a. The simplest model that can be fitted to the data is preferred
    b. The simplest model that explains the data is preferred [CORRECT]
    c. The most sophisticated model that can be fitted to the data is preferred
2. What are some alternatives to significance testing in online research?
    a. Effect sizes, parsimony criteria, and data visualization [CORRECT]
    b. Hierarchical regression, effect sizes, and parsimony criteria
    c. Hierarchical regression, effect sizes, and data visualization
3. What is overfitting?
    a. When a researcher applies too many models to the data

b. When a researcher uses a model that does not fit the current data perfectly

c. When a researcher uses a model that fits the data too perfectly [CORRECT]

Further Reading

- LaPlume, A. A., Anderson, N. D., McKetton, L., Levine, B., & Troyer, A. K. (2021). When I'm 64: Age -related variability in over 40,000 online cognitive test takers. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences.* Advance online publication. https://doi.org/10.1093/geronb/gbab143
- Troyer, A. K., Rowe, G., Murphy, K. J., Levine, B., Leach, L., & Hasher, L. (2014). Development and evaluation of a self-administered on-line test of memory and attention for middle-aged and older adults. *Frontiers in Aging Neuroscience*, *6*. https://doi.org/10.3389/fnagi.2014.00335

Web Resources

- Online cognitive assessment website: www.cogniciti.com
- R tidyverse suite for data manipulation: https://www.tidyverse.org/
- R package for data trimming: https://CRAN.R-project.org/package=trimr
- R coding cheat-sheets for data manipulation and visualization: https://www.rstudio.com/resources/cheatsheets/
- R data visualization resource: https://www.r-graph-gallery.com/

Acknowledgements

Drs. Nicole Anderson, Angela Troyer, and Brian Levine are thanked for helpful discussions and for supervising me on the design, methods, analyses, and publication of the original research study. Drs. Nicole Anderson and Angela Troyer are also thanked for insightful feedback on an earlier version of this case study.

References

Bender, A. R., Naveh-Benjamin, M., & Raz, N. (2010). Associative deficit in recognition memory in a lifespan sample of healthy adults. *Psychology and Aging*, *25*(4), 940–948. https://doi.org/10.1037/a0020595

Borella, E., Carretti, B., & De Beni, R. (2008). Working memory and inhibition across the adult life-span. *Acta Psychologica*, *128*(1), 33–44. https://doi.org/10.1016/j.actpsy.2007.09.008

Chiappe, P., Siegel, L. S., & Hasher, L. (2000). Working memory, inhibitory control, and reading disability. *Memory & Cognition*, *28*(1), 8–17. https://doi.org/10.3758/BF03211570

Buchanan, E.M., & Scofield, J.E. (2018). Methods to detect low quality data and its implication for psychological research. *Behaviour Research Methods*, *50*, 2586–2596. https://doi.org/10.3758/s13428-018-1035-6

Chetverikov, A., & Upravitelev, P. (2016). Online versus offline: The Web as a medium for response time data collection. *Behavior Research Methods*, *48*, 1086–1099. https://doi.org/10.3758/s13428-015-0632-x

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596–610.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods, 48,* 1–12. https://doi.org/10.3758/s13428-015-0567-2

Dykiert, D., Der, G., Starr, J. M., & Deary, I. J. (2012). Age differences in intra-individual variability in simple and choice reaction time: systematic review and meta-analysis. *PloS One*, *7*(10), e45759. https://doi.org/10.1371/journal.pone.0045759

Farhan, S. M., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., Greenberg, B., Grimes, D. A., Hegele, R. A., Hudson, C., & Kleinstiver, P. W. (2017). The Ontario neurodegenerative disease research initiative (ONDRI). *Canadian Journal of Neurological Sciences*, *44*(2), 196-202. https://doi.org/10.1017/cjn.2016.415

Feenstra, H. E. M., Vermeulen, I. E., Murre, J. M. J., & Schagen, S. B. (2017). Online cognition: Factors facilitating reliable online neuropsychological test results. The Clinical *Neuropsychologist*, *31*(1), 59–84. https://doi.org/10.1080/13854046.2016.1190405

Germine, L., Reinecke, K., & Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist*, *33*(2), 271–286. https://doi.org/10.1080/13854046.2018.1535662

Grange, J. A. (2015). *trimr: An implementation of common response time trimming methods.* R package version 1.0.1. https://CRAN.R-project.org/package=trimr

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science, 26*(4), 433–443. https://doi.org/10.1177/0956797614567339

Hilbig, B. E. (2016). Reaction time effects in lab-versus web-based research: Experimental evidence. *Behavior Research Methods, 48,* 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Hosmer Jr, D. W., Lemeshow, S., & May, S. (2011). *Applied survival analysis: regression modeling of time-to-event data* (Vol. 618). John Wiley & Sons.

Hultsch, D. F., MacDonald, S. W., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(2), P101-P115. https://doi.org/10.1093/geronb/57.2.P101

Kilkenny, M. F., & Robinson, K. M. (2018). Data quality:“Garbage in–garbage out”. Health Information Management Journal, 47(3), 103-105. https://doi.org/10.1177/1833358318774357

Kim, Y., Huang, J., & Emery, S. (2016). Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of medical Internet research*, *18*(2), e41. doi:10.2196/jmir.4738

Miller, J. B., & Barr, W. B. (2017). The technology crisis in neuropsychology. *Archives of Clinical Neuropsychology*, *32*(5), 541–554. https://doi.org/10.1093/arclin/acx050

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40-48. https://doi.org/10.1037//0021-843x.110.1.40

Myerson, J., Emery, L., White, D. A., & Hale, S. (2003). Effects of age, domain, and processing demands on memory span: Evidence for differential decline. *Aging, Neuropsychology, and Cognition*, *10*(1), 20–27. https://doi.org/10.1076/anec.10.1.20.13454

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging*, *23*(1), 104–118. https://doi.org/10.1037/0882-7974.23.1.104

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, *17*(2), 299–320. https://doi.org/10.1037/0882-7974.17.2.299

Parsons, T. D., McMahan, T., & Kane, R. (2018). Practice parameters facilitating adoption of advanced technologies for enhancing neuropsychological assessment paradigms. *The Clinical Neuropsychologist*, *32*(1), 16–41. https://doi.org/10.1080/13854046.2017.1337932

Paterson, T., Sivajohan, B., Gardner, S., Binns, M. A., Stokes, K. A., Freedman, M., Levine, B., & Troyer, A. K. (2021). Accuracy of a self-administered online cognitive assessment in detecting amnestic mild cognitive impairment. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, gbab097. Advance online publication. https://doi.org/10.1093/geronb/gbab097

Periáñez, J., Rioslago, M., Rodriguez-sanchez, J., Adroverroig, D., Sanchez-cubillo, I., Crespofacorro, B., Quemada, J., & Barcelo, F. (2007). Trail making test in traumatic brain injury, schizophrenia, and normal ageing: Sample comparisons and normative data. *Archives of Clinical Neuropsychology*, *22*(4), 433–447. https://doi.org/10.1016/j.acn.2007.01.022

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods, 39,* 365–370. https://doi.org/10.3758/BF03193004

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods, 47,* 309–327. https://doi.org/10.3758/s13428-014-0471-1

Rey-Mermet, A., & Gade, M. (2018). Inhibition in aging: What is preserved? What declines? A meta-analysis. *Psychonomic Bulletin & Review*, *25*(5), 1695–1716. https://doi.org/10.3758/s13423-017-1384-7

Rose, L. T., & Fischer, K. W. (2011). Garbage in, garbage out: Having useful data is everything. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 222-226. https://doi.org/10.1080/15366367.2011.632338

Salthouse, Timothy A. (2017). Shared and unique influences on age-related cognitive change. *Neuropsychology*, *31*(1), 11–19. https://doi.org/10.1037/neu0000330

Salthouse, Timothy A., Toth, J., Daniels, K., Parks, C., Pak, R., Wolbrette, M., & Hocking, K. J. (2000). Effects of aging on efficiency of task switching in a variant of the Trail Making test. *Neuropsychology*, *14*(1), 102–111. https://doi.org/10.1037/0894-4105.14.1.102

Slote, J., & Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods, 48,* 553–566. https://doi.org/10.3758/s13428-015-0599-7

Sunderland, K. M., Beaton, D., Fraser, J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., Peltsch, A. J., Pieruccini-Faria, F., Sahlas, D. J., Swartz, R. H., ONDRI Investigators, Strother, S., & Binns, M. A. (2019). The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project. *BMC medical research methodology*, *19*(1), 1-16. https://doi.org/10.1186/s12874-019-0737-5

Tombaugh, T. (2004). Trail making test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, *19*(2), 203–214. https://doi.org/10.1016/S0887-6177(03)00039-8

Tweedie, R. L., Mengersen, K. L., & Eccleston, J. A. (1994). Garbage in, garbage out: can statisticians quantify the effects of poor data?. *Chance*, *7*(2), 20-27. https://doi.org/10.1080/09332480.1994.11882492

Van Selst, M. & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. Quarterly Journal of Experimental Psychology, 47 (A), 631-650. https://doi.org/10.1080/14640749408401131

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Muller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686