Received: 1 November 2021 | Revised: 12 March 2022 | Accepted: 16 March 2022

DOI: 10.1002/icd.2315

REVIEW ARTICLE

WILEY

Adapting open science and pre-registration to longitudinal research

Isaac T. Petersen[1] | Keith S. Apfelbaum[1] | Bob McMurray[1,2,3]

[1]Department of Psychological and Brain Sciences, University of Iowa, Iowa City, Iowa, USA

[2]Department of Communication Sciences and Disorders, University of Iowa, Iowa City, Iowa, USA

[3]Department of Linguistics, University of Iowa, Iowa City, Iowa, USA

Correspondence
Isaac T. Petersen, Department of Psychological and Brain Sciences, University of Iowa, 175 Psychological and Brain Sciences Building, Iowa City, IA 52242, USA.
Email: isaac-t-petersen@uiowa.edu

Funding information
Eunice Kennedy Shriver National Institute of Child Health and Human Development, Grant/Award Number: HD098235; National Center for Advancing Translational Sciences, Grant/Award Number: UL1TR002537; National Institute on Deafness and Other Communication Disorders, Grant/Award Number: DC008089

Abstract

Open science practices, such as pre-registration and data sharing, increase transparency and may improve the replicability of developmental science. However, developmental science has lagged behind other fields in implementing open science practices. This lag may arise from unique challenges and considerations of longitudinal research. In this paper, preliminary guidelines are provided for adapting open science practices to longitudinal research to facilitate researchers' use of these practices. The guidelines propose a serial and modular approach to registration that includes an initial pre-registration of the methods and focal hypotheses of the longitudinal study, along with subsequent pre- or co-registered questions, hypotheses, and analysis plans associated with specific papers. Researchers are encouraged to share their research materials and relevant data with associated papers and to report sufficient information for replicability. In addition, there should be careful consideration of requirements regarding the timing of data sharing, to avoid disincentivizing longitudinal research.

KEYWORDS
development, longitudinal, open science, pre-registration, replication, reproducibility

Highlights

- Longitudinal studies have unique considerations that present challenges to standard models of open science.

- We propose a serial and modular approach to the registration of longitudinal studies; we emphasize the benefits of transparency and sharing data and materials.
- Increased adoption of open science practices may improve the rigor and replicability of developmental science.

In the last decade, psychology and social/biomedical science have wrestled with questions of scientific rigor and reproducibility. This was catalyzed by failed attempts to replicate (Camerer et al., 2018; R. A. Klein, Vianello, et al., 2018; Open Science Collaboration, 2015) or reproduce findings (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018), scientific misconduct (Craig et al., 2020; Stroebe et al., 2012), a growing appreciation of how analytic flexibility inflates evidence for scientific claims (Bakker et al., 2012; Simmons et al., 2011), and an understanding of how publishing biases intersect with these factors (Bishop, 2019; Ioannidis, 2005; Munafò et al., 2017). This has led to a crisis of confidence among both psychologists and the public (Gelman, 2016; Resnick, 2018, 2021), often referred to as the *replication crisis*. However, these issues are larger than failed replications and speak to a broader need to enhance rigor.

The field has converged on several aspects of research design and researcher behavior as the culprit for the crisis, some of which are particularly challenging in developmental work: (a) the extensive use of small samples (Button et al., 2013), which is common with difficult-to-recruit populations including children, (b) excessive freedom in analyses and methods, which is enhanced by large-scale studies with multiple covariates (Simmons et al., 2011), (c) insufficient reporting (Brown et al., 2014; Errington et al., 2021), (d) lack of standardization (Frank et al., 2017), (e) publication biases that suppress non-significant results, that is, the file drawer problem (Franco et al., 2014; Rosenthal, 1979), (f) statistical reporting errors (Bakker & Wicherts, 2011; Nuijten et al., 2016), (g) measurement error (Clayson et al., 2019; Loken & Gelman, 2017), (h) questionable research practices including selective reporting and Hypothesizing After Results are Known (HARKing), which can inflate evidence for exploratory work (John et al., 2012), and in some cases (i) outright fakery (Craig et al., 2020).

These are not the only problems (Oberauer & Lewandowsky, 2019), but many of these problems can be addressed by greater transparency about methods and analyses. Although the open science movement starts with openness and sharing, it is bound up in a broader conception of methodological rigor. Open science practices are thought to achieve this rigor in three ways. First, these practices make it easier to check each other's work by repeating analyses under different assumptions, or by facilitating replication experiments. Second, expectations of openness lead researchers to raise their methodological standards, knowing that others can check their work closely. Third, questionable practices like HARKing can be arrested by being open about hypotheses and when they were developed.

Open science practices are now common, and in some situations, mandatory. These include the following: pre-registration of hypotheses to prevent HARKing; pre-registration of methods and analyses to minimize researcher degrees of freedom; posting experimental materials, and videos of lab setup, procedures, and participants' behaviors on public repositories (Adolph et al., 2012; Gilmore & Adolph, 2017); mandating power analyses as a condition of publication; and sharing raw data and analysis scripts. These practices encourage more robust science, improve clarity about hypotheses and analyses, and enable meta-analyses and new insights (Gilmore, 2016).

Typical open science practices are straightforward and effective for clinical trials and small-scale cognitive and social psychology experiments. However, developmental psychology—and especially longitudinal studies—presents unique challenges to standard practice. This paper discusses some of these challenges, and it proposes guidelines for adapting open science to longitudinal research. These guidelines are not intended as authoritative. Rather, our

proposed guidelines are intended to spur discussion and iterative improvements that advance the mission of open science for longitudinal practices. We start by contrasting the structure of a typical longitudinal design with the kinds of studies for which standard open science practices were intended. Next, we describe the challenges of applying various opens science practices in longitudinal contexts. Finally, we offer our adaptation of these approaches and address open questions about this enterprise.

1 | LONGITUDINAL RESEARCH AND OPEN SCIENCE

Most open science practices are straightforward for cognitive and social psychology experiments with adults, and for cross-sectional and single-age experiments with infants and children. In such studies, data are collected at once or in a small number of sessions, and data are analyzed in a single wave to test a small number of hypotheses. In this situation, it is easy to determine analyses in advance and once the question is answered, there is no reason not to share the data. However, longitudinal studies do not fit this mold so easily.

Longitudinal studies track participants over time, sometimes for years. Long-term tracking of participants may have several consequences. First, the measures are often not completely known at the study's outset and may change as the study progresses. For example, in a study of language development spanning 7–12-year-olds, the appropriate measure of non-verbal cognitive skills for 11-year-olds might not be identified until Year 4 of the project; or in Year 2, the researchers might discover that a measure intended to be used throughout the project is too easy for older children. If the measure at a given age is too easy (ceiling effects) or too hard (floor effects), the measure's scores may show a restricted range, which makes it difficult to estimate individual differences, and attenuates associations with other variables. Second, because the study spans an extensive time-period, hypotheses will evolve as new findings arise, from the first years of the study, or from other labs. Finally, there might never be a clear point when the study is 'complete', because the complex data set allows evolving analyses long after data collection is complete, and grant renewals can extend the longitudinal window of data collection.

In addition, many longitudinal studies use large numbers of measures. This can be for several reasons. Latent constructs like language ability might be assessed with several measures to achieve greater validity. The use of multiple measures affords flexibility in combining these measures into indices, and this may in part be driven by properties of the data, such as the correlational structure and whether a class of models can be fit. Moreover, the scale and investment—human and financial—of a longitudinal study makes it naïve to plan a study around a single hypothesis. Rather, studies are often designed around a few core hypotheses but include a range of other measures and rich background and demographic data that provide grist for future exploratory work testing moderators of original hypotheses, or other hypotheses entirely.

Box 1 describes the design of two studies currently underway by the authors of this paper illustrating these features, including a description of the focal research questions, pre-registered hypotheses, and when and how

BOX 1 How we have adapted open science for two longitudinal studies

School readiness study: The *School Readiness Study* (current $N = 108$) is an ongoing NIH-funded longitudinal study of the development of school readiness during the transition from preschool to school entry. The project takes a bio-psycho-social perspective to understanding the processes that influence children's academic and behavioral readiness for school. Several hypotheses were pre-defined at study outset, including that smaller amplitudes in cognitive control-related event-related potentials (ERPs), including the N2 and P3 will be associated with externalizing problems and poorer academic skills, and that self-regulation will partially mediate these associations. ERPs are neuro-electrical responses to stimuli measured on the scalp

with electroencephalography. The N2 and P3 ERPs reflect the second negative deflection and the third positive deflection, respectively, in the brainwave in response to a stimulus. The N2 and P3 are of interest because of their relation to cognitive control. In addition, we also included sufficient additional measures to conduct exploratory work on factors like child temperament and parent self-regulation.

The study uses an accelerated longitudinal design. Children were recruited beginning in 2018 at 36, 45, 54, or 63 months of age, and are assessed every 9 months over 4 time points. Thus, any given child is followed for 2¼ years and the final sample spans 3 to 7.5 years of age. We anticipate cross-sectional papers early in the project, followed by longitudinal papers later. This raises the question of when it is appropriate to share data, and what should be shared at each point.

The assessment battery assesses the core questions using multiple measures of self-regulation, including neural tools (like ERPs) that require complex time series analyses to be useful, and cognitive tasks assessing cognitive control and attention. Time series analyses involve the analysis of high-density data in the temporal domain, for example, temporal principal component analysis. In this study, the data are nested across multiple levels: multiple timepoints of EEG data (level 1) are sampled from the same person at a given measurement occasion (level 2), and the longitudinal design yields multiple measurement occasions over time within the same participant (level 3). The sophistication of these data and approaches can make it challenging to pre-define indices of cognitive control in advance. This project combines these measures with multiple parent- and teacher-reported questionnaires that assess emotional and behavioral regulation, behavior problems, academic skills, and environmental factors. For these measures, we anticipate combining related tools using factor analyses, but the number of factors cannot be known prior to analysis, making pre-registration of statistical analyses challenging. In addition, we made some changes to the measures in the longitudinal study based on issues such as range restriction. For instance, we replaced an emotion regulation task at T2 because it did not elicit sufficient individual differences. We also stopped assessing one of the perceptual inhibitory control tasks after 63 months of age due to ceiling effects.

As a result of these challenges and changes, the *School Readiness Study* uses a version of the modular registration approach proposed here. Hypotheses, methods, measures, and analyses are registered separately both before and during data collection. The focal hypotheses, methods, and measures were preregistered on the project's main Open Science Framework (OSF) page: https://osf.io/jzxb8. Changes to registrations are timestamped on the OSF. Each paper associated with the project will have its own OSF page, including the data (i.e., variables) used for that paper, a data dictionary for those variables, the analysis code, and a computational notebook.

Growing words project: The *Growing Words* Project ($N = 242$) is an ongoing NIH-funded longitudinal study of the development of spoken- and written-word recognition during the school-age years. The project uses an accelerated longitudinal design that began testing 1st, 2nd, and 3rd graders in the spring of 2021, and will test these cohorts yearly for 4 years. In particular, *Growing Words* uses eye-tracking in the Visual World Paradigm (Rigler et al., 2015) to examine the real-time mechanisms of word recognition. This is related to standardized outcome measures of language and reading, to experimental measures of cognitive control and to structural magnetic resonance imaging (MRI).

These data are used to investigate a number of hypotheses. First, we ask whether real-time processing skills develop from earlier achievements in language ability, and/or if real-time skills enable better language. Second, we investigate the impact of cognitive control on word recognition. Finally, we seek to understand the role of reading development in changing word recognition and speech perception. We also include exploratory measures of the language/literacy home environment and phonological processing (to name a few) which did not have explicit hypotheses detailed at study outset.

The sophisticated eye-tracking measures will need to be collapsed into index measures (see Box 2). There are several unknowns, and there is insufficient current data on reliability with children. As a result,

we anticipate using some pre-defined indices (that have 'worked' in prior smaller scale studies), as well as database measures to define indices. Critically, these indices will be identified in Year 1 and used for all papers on these data.

Growing Words completed Year 1 in 2021, and as of 2022 we are already anticipating changes to the study. At least one language measure may be replaced; we are debating subtle changes to the eye-tracking measures to improve their measurement properties based on Year 1 data collection; and we have identified one small issue with one of the cognitive control tasks that will be fixed. Although initial methods are now registered, these changes and their motivations will be documented in separate registrations.

The *Growing Words* project will generate very large datasets. Individual eye-tracking assessments (there are five) may each generate 10–20 gigabytes of raw data making it difficult to share these directly. Instead, we will release indices used in specific papers as the papers are published. Raw data will be available upon request.

data/materials will be shared. These features of longitudinal studies make it challenging to apply standard tools of open science to longitudinal studies.

2 | TOOLS FOR OPEN SCIENCE

Open science consists of a loose bundle of practices: pre-registration, data and materials sharing, power analyses, rigorously guarding against *p*-hacking, and full disclosure of analytic choices. These can be independent of one another—data can be shared regardless of whether the design was pre-registered. Few labs adopt them all. Rather, practices are assembled based on the needs and nature of each study. Although several of these practices, including power analyses and materials sharing, are straightforward to apply to longitudinal projects, two are particularly difficult: pre-registration and data sharing. These are the focus of this paper.

3 | PRE-REGISTRATION

3.1 | Standard approaches to pre-registration

Pre-registration involves publicly posting a study's design, hypotheses, methods, materials, and analysis plan before data collection begins. The pre-registration plan may be submitted to a journal as a registered report[1] and reviewed before data collection—see Table 1 for developmental journals that accept registered reports. More typically, pre-registration involves simply posting this information in a permanent repository. Some federal grants, such as NIH-defined clinical trials, require pre-registration to reduce publication bias and to ensure that negative results from clinical trials are released and publicly known. Making methodological and analysis decisions before data collection prevents design, analysis, or interpretive decisions that are contingent on the desired result, and it ensures that hypotheses are specified in advance and not changed after examining the data.

A research project has many decision points (Frank et al., 2017), including (a) what counts as 'piloting', (b) when it is acceptable to restart a testing session due to technical difficulties or participant non-compliance, (c) which numerical indexes to use in analysis, (d) whether and how to exclude data or remove outliers, (e) whether and how to transform the data, (f) how to analyze the data, and many others. These decisions are sometimes referred to as *researcher degrees of freedom* (Simmons et al., 2011), and can lead to widely varying outcomes (Gelman &

TABLE 1 Developmental journals that accept registered reports

British Journal of Developmental Psychology
Developmental Cognitive Neuroscience
Developmental Science
Infancy
Infant Behavior and Development
Infant and Child Development
Journal of the American Academy of Child and Adolescent Psychiatry
Journal of Child Psychology and Psychiatry Advances
Journal of Cognition and Development

Loken, 2013). Monte Carlo simulations have shown that when these decisions are made after looking at the data, Type I error can inflate to substantially greater than 0.05 without the researcher or reader being aware of it (Simmons et al., 2011). Pre-registration reduces researcher degrees of freedom and limits reporting biases, such as reporting only significant results or only particular study conditions, to keep the Type I error rates at the intended alpha.

Pre-registration is commonly used for hypotheses, methods, and statistical analysis, but can also be used to document the rationale for these decisions. Pre-registration is valuable for planning and thinking through how hypotheses map to patterns of results. Pre-registration may also constrain statistical analyses because many approaches to family-wise error correction differ based on whether research questions are confirmatory—that is, hypothesis-driven—or exploratory (Bender & Lange, 2001; de Groot, 2014; Francis & Thunell, 2021; Rubin, 2017). Even if the complete statistical model cannot be known in advance, pre-registering hypotheses allows the researcher to determine which approach is most appropriate; waiting to designate a test as such until after results are known can lead researchers to take inappropriate liberties.

Pre-registration is not a panacea. Pre-registration was designed for studies that test: (a) a small number of hypotheses with (b) well-defined methods and (c) analytic models that can be clearly designed in advance. Canonically, this is done on a well-developed paradigm that has seen extensive prior testing, often which is not pre-registered. Consequently, pre-registration does not lend itself as easily to studies using less-established methodologies, where measures or indices such as composite scores may not be fully established, or the variance and covariance structure are not well enough understood to specify a model. Perhaps most challengingly, the lengthy time-period and high cost of a longitudinal study make it unlikely for there to be extensive prior longitudinal studies on the same topic with the same measures, again making it difficult to make these decisions. Finally, an over-reliance on pre-registration can lead researchers to miss important findings that were not hypothesized before data collection. This is particularly likely when the study lasts years.

3.2 | Pre-registration in longitudinal designs

Standard pre-registration models work well for cross-sectional studies or small single-visit studies that answer a single question. In this context, pre-registration has driven important advances and consortia (Frank et al., 2017; Gilmore, 2016; Gilmore & Adolph, 2017). However, pre-registration is more challenging in individual-differences studies, longitudinal work, and research with unique populations. This is for several reasons.

First, at the outset of a longitudinal study, researchers are unlikely to anticipate all research questions and hypotheses that will be generated over the course of the study. Because longitudinal studies are time-extended and multidimensional in nature, hypotheses may be developed and refined while data collection is underway, in response

to new findings in the data, observations of participants in the study, emerging findings from other labs, novel methods, or time for ideas to percolate. This blurs the line between confirmatory and exploratory work, which can be disqualifying in many scientific sub-cultures and some journals that expect clear specification of whether a study is confirmatory or exploratory or do not place significant value on exploratory or descriptive work. Moreover, not only may hypotheses change over the course of a study, but so might methods.

Such changes may be aligned with the goals of the study but perhaps not anticipated. For example, a measure used in Year 1 may prove to have poor psychometric properties and is changed in Year 2. This could occur if older children unexpectedly reach the ceiling, or an unexpected practice effect emerges from using the measure at multiple testing sessions. Other times, this may represent a deeper change. Some research questions and hypotheses may depend on empirical or analytic methods that do not exist at the study outset. Quantitative or empirical methods may also become available after the study begins that motivate data re-processing or re-analysis to answer new questions. For instance, recent algorithms for assessing a person's heart rate from video (Hassan et al., 2017) could allow re-analysis of older video recordings.

Experimental protocols may also need to be refined during longitudinal data collection in response to unexpected contingencies that result in non-random data loss. For example, the authors of this paper are conducting a longitudinal study (the *Growing Words* Project) that includes several measures of language, reading, and cognition. In Year 1, some low-performing students took longer than anticipated to complete some tasks, which led them to be more likely to not complete the final task in the session. This led to a bias in data loss among poorer performers even though they did not have difficulty with that specific task. This necessitated a change in protocol to ensure that all groups were equally likely to complete that task.

Some variables in longitudinal designs may also be difficult to pre-define. A construct like a language or cognitive control is often represented by a latent factor constructed across several tasks. Critically, measures' factor loadings may change over time. This may even be true within a measure—items that indicate an externalizing disorder at 4 years of age (e.g., 'throws tantrums') may be irrelevant at 18, and vice versa ('uses illegal drugs') due to changes in the construct's manifestation (Petersen et al., 2020). Even if the researcher has hypotheses regarding the factor structure of a set of variables, the number of factors across a set of measures, and the loadings of the variables within each factor at each age cannot always be anticipated in advance.

This can get more complex for longitudinal designs that employ sophisticated measures like eye-tracking (Law et al., 2017), electroencephalography (EEG; Bell & Cuevas, 2012; Brooker et al., 2020), or pupillometry (Hepach & Westermann, 2016; Winn et al., 2018). These dense measures are often collapsed into single indices that reflect simple constructs (e.g., rate of activating a word: c.f., McMurray et al., in press)—see Box 2. This conversion from a dense time series to a single index may be difficult without looking at the data—for example, specifying timepoints or understanding the shape of the curve for a fitting procedure. Consequently, it may not be clear at the study outset, which indices best assess a construct at each timepoint.

Registration of the statistical plan is the next step. This can be done if indices are known in advance, or indices can be left as placeholders. However, in longitudinal work, the planned statistical approach may not be appropriate once data are collected. For example, a mixed model may not fit the data—for example, if the distribution of residuals is unexpectedly not Gaussian or a proposed random slope does not converge. Consequently, researchers may need a different model or a different approach. Other times, findings may lend themselves to a better approach. For example, Fellman et al. (2020), present a working memory training study that was originally pre-registered with a standard null hypothesis test; however, the effects were small, consistent with other concurrent studies, leading them to switch to a Bayesian approach to provide more insight. We commend these authors for their openness about this process. However, a rigid pre-registration could minimize the insight to be gained by a change in analysis.

In addition, the difficulty of pre-registration depends on the scope of the study. The larger the scale of the project, the more challenging it may be to pre-register. Large-scale longitudinal studies often have tight timelines and multiple collaborators, each of whom may have their own hypotheses, some of which may conflict with

BOX 2 Example of the many ways in which complex measures can be condensed into index variables

Complex measures often present multiple possible analysis approaches, which create potential researcher degrees of freedom. Often, these measures are condensed into summary measures to reflect simple constructs. There are various possible data reduction strategies for this. Choosing an analysis strategy offers a major researcher degree of freedom; an unscrupulous approach could attempt different ways to condense measures to see if any support the hypothesis, and then report only this approach. Here, we illustrate the example of condensing continuous eye-tracking data, which includes samples of looks at frequent intervals (e.g., every 4 ms) to different types of displayed objects (e.g., targets, competitors, unrelated objects), into index variables of relevant constructs.

We highlight this for one such construct that is commonly assessed for eye-tracking studies of word recognition: the speed of target recognition. Data are often plotted as in Figure B.1a. This figure plots preliminary data from a measure of spoken-word recognition in the *Growing Words* project for 1st and 3rd graders. It plots the proportion of looks to the target object at each 4 ms time point, averaged across trials and participants, and with separate curves for different grades. It seems apparent from these data that 3rd graders look to targets more rapidly than 1st graders. However, these data are comprised of many only partially independent data points, that have been averaged for this visualization. These data can be converted into a single measure of 'speed of target recognition' that can be statistically analyzed in several different ways. For example, parametric curves could be fit to the data, and the parameters of these curves

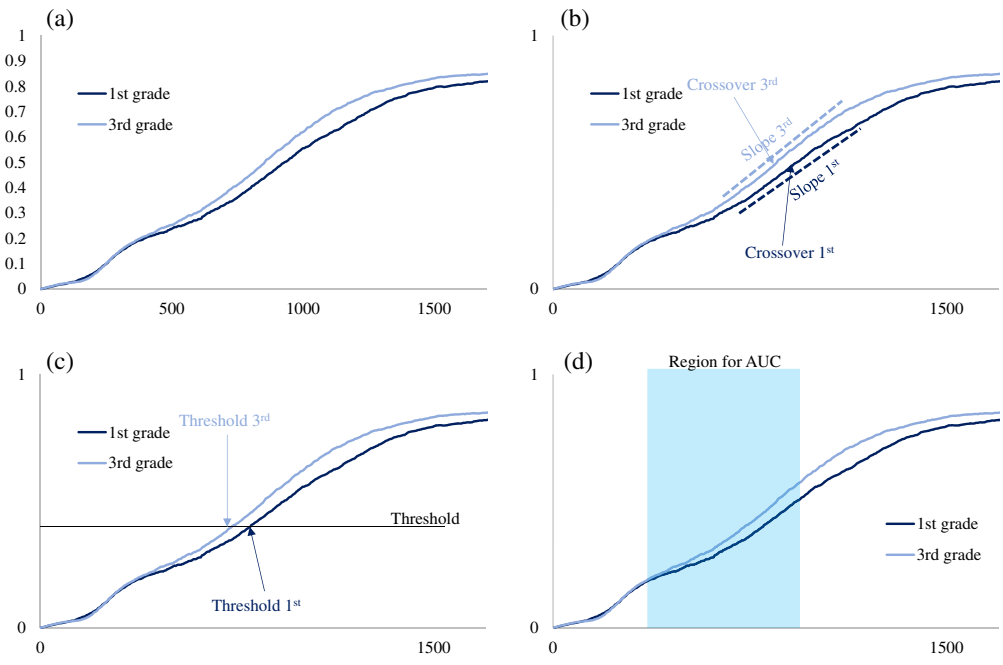


FIGURE B.1 Different approaches to operationalizing traget recognition speed. (a) Average proportion of looks to the target across time by grade. (b) A logistic curve can be fit to the data, and the slope and/or crossover parameters can index target fixation speed. (c) A threshold of looking can be set, and the time when this threshold is exceeded can be used to index target fixation speed. (d) A region of time can be set during which the area under the curve can index speed of target looking

used as the indices (Figure B.1b). In this case, the slope of the logistic function that describes the data could signify how rapidly target looks rise; however, we could also use the timing of the crossover point of this logistic function. A third option within this approach is to combine the slope and timing into a single score representing timing (McMurray et al., 2019). However, numerous other ways to operationalize this speed are also possible. One could assess the time when the curves first cross some threshold (e.g., when they first exceed 0.4; Figure B.1c). This approach creates additional decision points, including what threshold to use, and what counts as exceeding it (e.g., what to do if a participant exceeds it but then later drops back below the threshold). Yet another approach is to measure the area under the curve of looks to the target over some time window (Figure B.1d) to determine if one group shows more overall target looks than the other in this transition window; this approach again introduces further decisions of what time window to use. Countless other operationalizations are also possible—counting the number of saccades to the target within some time window, measuring entropy of eye movements to objects across time, etc.—each of which comes with its own set of analysis decisions.

Several approaches to operationalizing the construct of target recognition speed have merit, and different theoretical orientations may lend themselves to different decisions. Further, some measures may prove more reliable than others, and thus be more appropriate indices of the construct. The rationale behind the choice of the operationalization should be described in advance. This may take the form of commitment to a single approach—for example, if the data will be compared against prior studies that used a particular operationalization. Alternatively, it may take the form of a strategy for identifying the best operationalization—for example, detailing an approach to estimating reliability of measures, with a clear explanation of what metric will dictate the best measure. In either case, however, these plans should be described in advance of analysing the data. This approach will help prevent the various possible decisions in data reduction from leading to *p*-hacking.

An example of our pre-registration approach to measure identification for eye-tracking data is available at the Growing Words OSF site (https://osf.io/vzb2k).

each other or depend on hypotheses or findings from other domains of the project. Thus, flexibility may be necessary for the pre-registration of large longitudinal studies, and costs and time of pre-registration should be factored into the grant.

These challenges raise questions of whether it is appropriate to completely pre-register longitudinal designs. If researchers are restricted to only examining hypotheses specified in pre-registration, they could miss important findings or hypotheses that emerge over the study. Locking the methods could make it difficult to make a much-needed course correction, which could be catastrophic considering the scale and cost of many longitudinal studies. Pre-registration could constrain analysis decisions to inappropriate data and model structures.

One could just decide that all analyses of longitudinal designs are 'exploratory'. However, this neglects the benefits of pre-registration *even for exploratory research*. For example, researcher degrees of freedom can be huge in large longitudinal datasets, and these must be responsibly constrained. Further, longitudinal designs are rarely purely exploratory but are a hybrid in which some hypotheses are specified in advance, whereas other measures are included for exploratory purposes. In today's culture, 'exploratory' is a bit of a pejorative—though it should not be. Thus, deeming longitudinal work exploratory and skipping pre-registration undermines the value of longitudinal designs, which are the gold standard in developmental science for describing, predicting, and understanding change over time. Instead, the field needs to preserve openness and rigor of pre-registration, without mandating that all decisions be made prior to the study and that all decisions be made at once.

4 | DATA/MATERIALS SHARING

4.1 | Standard approaches

A second open science practice is sharing materials: manuals, protocols, consent forms, stimuli, lab notebooks, data, metadata, data processing syntax, statistical analysis code, detailed results, computational notebooks, and preprints. These are usually shared at the conclusion of the study, that is, when papers are submitted in an online repository, though some labs share materials incrementally, as the project unfolds. Sharing materials makes science more accessible. It allows other researchers to review a research study more comprehensively and identify potential errors in design, stimuli, or analyses, and reproduce, replicate, and extend findings. A particular benefit is error detection and correction.

Traditional, non-open science approaches share data and materials only at summary levels, through descriptions in published manuscripts or by direct request to the author. This presupposes that research materials are exactly as described, with all necessary information and without errors. However, it is challenging to fully describe years of labor and hundreds of lines of code in readable methods and results. This is further challenged by the increasingly short page limits of many journals and the fact that online supplements rarely receive the same level of review as the actual manuscript. Moreover, analyses of authors' response rates to requests for data suggest only low-to-moderate compliance (27%–59%; Tedersoo et al., 2021; Wicherts et al., 2006). Thus, summaries and personal requests are insufficient.

In contrast, complete access to the data—either before publication or as a condition of it—has numerous benefits. First, it keeps researchers statistically transparent because reviewers and colleagues have immediate access to the data. This can minimize questionable research practices (John et al., 2012), such as multiple testing, flexible use of covariates, and selective reporting, that is, '*p*-hacking', because others may detect this. Second, data sharing provides a second round of accuracy checking when data and code are commented and curated for posting. Finally, other researchers can use the data for other reasons. Alternative theoretical approaches might suggest different statistical models or new questions, people may reanalyze data for methodological purposes such as evaluating reliability, and they can be used in meta-analyses where direct re-analysis is preferred over published effect sizes.

Illustrating the value of this, data sharing funding and curation initiatives have made data and codebooks from large-scale longitudinal studies available for public or restricted use, such as the Adolescent Brain Cognitive Development (ABCD) study (Casey et al., 2018), the National Longitudinal Study of Adolescent to Adult Health (Add Health; Resnick et al., 1997), the Early Childhood Longitudinal Study (ECLS) Program (Tourangeau et al., 2009), the National Longitudinal Surveys (NLS; Chase-Lansdale et al., 1991), and the NICHD Study of Early Child Care and Youth Development (SECCYD; NICHD Early Child Care Research Network, 2005). The large number of secondary studies that have resulted from these projects attests to the value of longitudinal data sharing.

The benefits of material sharing are evident, and this practice improves research quality and confidence in findings. However, there are also challenges and potential pitfalls of sharing materials. The pitfalls may not be in sharing materials per se, but in the sole reliance on using materials that have been shared by others. First, when conducting replications using shared materials, any issues with the original materials risk recurring. This is particularly the case in subfields like language or perception, where designing stimulus materials is a significant undertaking; here, the ease of using shared materials may mean that inadvertent mistakes or irrelevant design choices (e.g., who the speaker was in a language experiment) get carried over from one study to the next. For example, Strand (2020) described a scenario where a coding error led to a spurious finding, which was replicated by another research team that used her materials. There is a risk of overconfidence in the fidelity of shared materials (or any materials, for that matter). In some instances, conceptual replications or replications with novel materials may be more compelling than direct replications. Although nothing in the open science paradigm prevents recurring mistakes in replication studies, the ease of accessing others' work may make a direct replication more attractive than a conceptual replication. Nevertheless, sharing data and research materials may accelerate the detection and correction of mistakes (Gilmore et al., 2021).

Second, what should be shared and when? Data sharing can solidify statistical findings and benefit the field. But it also means that data are now outside of the control of the originating lab, and other laboratories may publish results

based on it. This is particularly an issue with large longitudinal data sets consisting of many complex measures. Oftentimes, the person requesting the data signs an agreement or writes a statement about how they will use the data. Such steps may provide greater accountability for the appropriate and intended use of data collected by others. Moreover, publishing multiple times using the same data may lead to increased Type I errors if corrections are not made for multiple testing (Thompson et al., 2020), and may limit the unique contribution of a paper (Kirkman & Chen, 2011).

4.2 | Data sharing in longitudinal designs

Longitudinal designs raise two primary issues with data sharing: ethical obligations to participants and the issue of what to share and when to share it.

4.2.1 | Ethical concerns

One challenge with data sharing is the potential identifiability of participants (Gilmore et al., 2021). Standard approaches to de-identifying data can be sufficient for single-visit small-scale studies. However, longitudinal studies often have large quantities of data that could be stitched together to identify participants (Gilmore, 2016). It is the responsibility of the principal investigator to share data in a way that protects the identities and confidentiality of participants while ensuring the data are usable by others. However, secondary data users share the obligation of protecting participants (APA Data Sharing Working Group, 2015). This becomes even more important when the project deals with sensitive data, such as geographical data, biological data, illegal (e.g., substance use) or stigmatized behavior (e.g., same-sex intercourse), health conditions (e.g., mental disorders or sexually transmitted diseases), or qualitative data such as interview transcripts, including clinical interviews. This is a particular challenge for longitudinal work, which often includes a large array of measures and background information. When dealing with potentially identifiable or sensitive data, procedures for sharing data should be carefully monitored. Nevertheless, researchers may share raw, identifiable data if participants are properly informed of the risks to privacy and confidentiality, and if they provide consent for identifiable data to be shared (Gilmore et al., 2021; Gilmore & Qian, 2021; Meyer, 2018). For instance, the Databrary project (Simon et al., 2015) provides templates for obtaining consent for sharing identifiable data with researchers who have been authorized to access the data and who have signed an access agreement to use the data in accordance with ethical principles.

Additionally, there is an ongoing discussion among researchers and their institutional review boards (IRBs) as to whether participants who were minors during data collection need to be reconsented when they become adults (Berkman et al., 2018). This could occur if longitudinal data collection spans this transition, or if data from minor years are shared after participants have become adults. Many researchers find the idea of reconsenting minors when they become adults impractical and a barrier to sharing data (Gilmore, 2016).

4.2.2 | What to share and when to share it

The time-extended nature of data collection in longitudinal studies along with the large quantity of data collected makes it unclear what data to share and when. At a practical level, longitudinal studies involve large quantities of data that take lots of time to code and clean for analyses. Data sharing adds the burden of documentation to make the data interpretable by others. What may be a relatively simple task for a small experiment can become months of work in a longitudinal study. The difficulty of sharing data or materials may depend on the scope of the study. The larger the scale of the project, the more funding may be available for data curation and sharing. Even so, library scientists may be able to provide help in preparing data for sharing (Soska et al., 2021). Although sharing data can involve

considerable work, it is important to balance staff time and the taxpayers' investment in research against the prospective benefits of sharing a particular dataset at a particular time.

Even beyond the scale of data, the time-extended nature of longitudinal projects presents challenges. Some analyses may not be able to proceed until data collection at all measurement points is complete (Eisenberg, 2015). Longitudinal studies are rarely designed to answer a single question but also include many measures for secondary questions or for exploratory work for later studies. This raises several concerns for data sharing. First, a reasonable concern is that others may publish based on the shared data before the original investigators can, that is, 'getting scooped' (O. Klein, Hardwicke, et al., 2018). This concern is particularly relevant for longitudinal studies. The protracted data collection and the rich set of measures in a longitudinal study may permit others to push ahead on a new question—even one planned by the study team—while the study team is working on a different question. Longitudinal studies require substantial personnel, time, energy, and money; longitudinal studies would be disincentivized if researchers were required to share all data fully and publicly as they are collected or as soon as data collection is complete (Eisenberg, 2015). Others have argued that the worry of being scooped is unwarranted in psychology because (a) most subfields are not so competitive that they are populated by researchers who are racing to publish a particular finding, and (b) the benefits of increased exposure outweigh the possibility of being scooped (O. Klein, Hardwicke, et al., 2018). Nonetheless, researchers' concerns about this problem—whether warranted or not—likely shape the decision to invest the effort in data sharing.

There are some solutions to concerns about getting scooped. Data can be embargoed for a certain period, and journals could require those who use others' data, materials, or analysis code to cite the original source—most repositories provide persistent identifiers (DOIs) that can be cited (Gennetian et al., 2020). Nonetheless, these may not be sufficient for the complexities of longitudinal work. Given the importance of longitudinal designs for answering questions about developmental mechanisms, data sharing procedures must incentivize longitudinal designs by giving the lead researchers the first 'crack at' key analyses (Eisenberg, 2015).

In addition to the issue of when to share is the issue of what to share. Sharing need not be all or none. Researchers may choose to share data from only a subset of measures but not data from other measures, such as sensitive data from a clinical assessment or measures that have not been analyzed yet. Moreover, indices of core constructs may be derived indices—latent factors, or summary statistics from complex measures like eye-tracking or event-related potentials. What should be shared? Indices? Individual item responses for each participant? Each 4 ms of eye-movement data? Although some would argue to share everything, there is a cost to the investigator for assembling these data—both in terms of time and the possibility of being scooped. There may also be costs to the end user (and the ultimate impact of sharing the data), if data are shared in the wrong form: data that are too 'raw' may require extensive rescoring, and processing. Moreover, raw data may not be as useful for things like meta-analyses or for people not deeply versed in a given instrument. Thus, what is appropriate and useful to share is not always straightforward.

Another potential barrier to sharing data and materials is that others may discover errors, and researchers may fear having their reputation or abilities publicly undermined (Gilmore, 2016; O. Klein, Hardwicke, et al., 2018). This concern may be particularly acute for longitudinal designs whose scale creates many opportunities for minor errors. To combat this, shifts in scientific culture may be necessary to promote openness. We join calls to reduce the blame on others whose work fails to replicate (Gilmore, 2016), and we encourage researchers to keep scientific critiques focused on methods, not the person—whether in journals, conferences, lab meetings, or social media. In addition, journals may consider alternatives to retractions: publishing revised and corrected versions of retracted articles, that is, retraction with replacement (Strand, 2020).

5 | GUIDELINES FOR ADAPTING OPEN SCIENCE TO LONGITUDINAL STUDIES

Developmental science has lagged behind cognitive and social psychology in implementing open science practices (Frank et al., 2017). Some of this lag likely arises from longitudinal designs which do not lend themselves easily to

one-size-fits-all practices. Instead, there is a need for more flexible approaches that embrace goals of rigor, reproducibility, and openness, while accommodating longitudinal paradigms, which are systematic and clear enough to be implemented into pre-registration and open science platforms.

Open science practices are time-consuming to implement, challenging to navigate, and complex for longitudinal designs. Nevertheless, the benefits outweigh the costs, and every additional step toward transparency helps. The critical question is how best to adapt these practices for longitudinal work. The values of open science we emphasize here are intended to maintain rigor, reproducibility, and replicability. These include openness and transparency, constraining researcher degrees of freedom, pre-specified hypotheses and honesty regarding whether analyses are confirmatory or exploratory, and maintaining incentives to collect important longitudinal data. Indeed, in longitudinal research, data sharing may be especially important because longitudinal studies cannot be readily replicated (Adolph et al., 2012).

Our team—which is currently managing two longitudinal projects—has begun a series of discussions with statisticians and collaborators to develop such an approach (for discussions related to assessment and clinical science, see Tackett, Brandes, King, & Markon, 2019; Tackett, Brandes, & Reardon, 2019; Tackett et al., 2017). Our goals are to develop tractable procedures that increase the adoption of these open science practices in developmental science, and that maximize the benefits of open science while minimizing costs and challenges. We argue that the principal goal of any practice must be openness in reporting, and this is more important than whether this occurs before, during, and/or after the study. Reporting all methods and measures is challenging when publishing papers from longitudinal studies—the current zeitgeist is toward short papers that address a single question. Because papers may not fully report the complete set of measures in a longitudinal study, cherry-picking—selectively reporting measures that support a hypothesis—is a real possibility.

We propose a modular form of registration, that preserves openness, which separates methods, hypotheses, measures, indices (e.g., scores derived from multiple measures), and analyses. Each component may have its own timelines, and registration of each can occur at multiple points during the study. This more flexible approach offers multiple paths. Plans for specific components can be presented in advance but updated over the course of longitudinal data collection. For other components, plans can be developed after data collection begins, as long as openness is maintained about when and by whom decisions were made, and what information guided decision making. For each component, a clear explanation of the rationale of forthcoming analyses and approaches supports the overall goal of transparency.

Similarly, we propose data sharing be linked to specific papers or phases of this pre-registration process. Releasing data in this piecemeal approach is valuable. This approach may not release all data simultaneously and in real-time, but it meets the goal of allowing others to replicate or extend analyses. We next detail our proposed approach, starting with what not to do.

5.1 | What not to do

John et al. (2012) provide a list of ethically questionable research practices to avoid. Two are particularly relevant for longitudinal studies.

First, researchers should not cherry-pick ages or measures from a broader longitudinal study for reporting, a form of *p*-hacking. When such decisions are made to simplify the analysis, they should ideally be made before analysis—though perhaps not before data collection—and should be transparently disclosed.

Second, researchers should not present exploratory work as confirmatory. However, HARKing comes in two forms. Secretly HARKing in the Introduction section (SHARKing) is problematic—though the confirmatory expectations of many reviewers and editors encourage it. By contrast, Transparently HARKing in the Discussion section (THARKing; Hollenbeck & Wright, 2017) can be valuable by providing post hoc explanations for surprising findings in the Discussion section, and testing those possibilities using transparent post hoc exploratory analyses (Hollenbeck & Wright, 2017). This is particularly valuable in rich longitudinal datasets.

5.2 | What to do

Longitudinal open science should be a gradual process that unfolds over time—much like a longitudinal project. The key principle is not that everything needs to be decided in advance, or immediately shared in its entirety. Rather, by modularizing the process, each stage of hypothesis generation, study design, planning, and data exploration and analysis, can be registered when it is ready and when it is available, and data sharing can follow a similar form. However, separating the pieces—both conceptually and in time—raises the expectations for what must be reported. Under this regimen, registration documents must go beyond simply reporting the plans: researchers must be transparent about when a decision was made, what information contributed to it, the rationale for that decision, and who made it—especially if there are diverging opinions on the team. If this expanded transparency is maintained, many benefits of open science can be realized without the enhanced costs of implementing them for a longitudinal project.

5.2.1 | (Not necessarily pre-) registration

Pre-registration fully in advance of data collection is not the only option. There is a continuum of approaches: (a) pre-registration, (b) *co-registration*—after data collection starts but before analysis—and (c) *post-registration*—after analysis has begun (Benning et al., 2019). We advocate for a piecewise co-registration—each incremental step toward transparency adds positive value (O. Klein, Hardwicke, et al., 2018). For longitudinal studies, a single registration approach may not always be optimal, as long as researchers are transparent about the decision process.

We propose a serial and modular form of registration with a wall between phases of research: hypotheses; methods; index identification; and statistical analyses. Each can be registered separately and asynchronously for different aspects of the study, and before, during, or after data collection. Critically, in this mode, registration needs to go beyond 'just the facts',—the typical pre-registration model—to honestly reflect the source and development of a hypothesis, method, or index. Such a document should be clear about:

1. What is being registered (e.g., a hypothesis, index, analysis, etc.), and what is not being registered, whether because it will be registered in the future (e.g., a future hypothesis, a measure to be developed, protocol changes), was registered elsewhere, or will not be registered.
2. When (relative to the timeline of the study) it was developed.
3. Why it was developed—for example, it was planned at the outset, or it was developed in response to exploratory analysis, or a new paper.
4. What information was known or unknown when the hypothesis or analysis was planned.
5. Who (in the study team or the broader collaboration) was involved.

All of this should be described alongside a traditional registration of what the study team proposes to do. Critically, these documents should preserve a timeline of the registration alongside the timeline of the study.

For a large longitudinal study, it may make sense to initially register the methods at the outset. Thereafter, the modular approach allows multiple streams. While registration of hypotheses, measures, indices, and analysis are contingent on each other, one could run this process in parallel for different aspects of the study. For example, in the *Growing Words* project, core hypotheses about speech perception could go through this registration process immediately, whereas the process of registering hypotheses, indices, and models for secondary questions about the language and literacy environment could start later.

Not all research needs to be confirmatory. Exploratory research is an important complement to confirmatory research (Barbot et al., 2020); indeed, exploratory work is perhaps the purest form of 'discovery', and rich longitudinal datasets present many opportunities for this. Post hoc data exploration may leverage the value of costly data (Barbot et al., 2020; Hollenbeck & Wright, 2017; Rosenthal, 1994). A modular approach can leverage this: for

example, researchers can register hypotheses or models when they are known but can leave other questions for exploratory work. Exploratory questions and analysis plans can also be registered separately. Thus, our registration model leverages the values of both.

In this modular form, we see several loci of registration—though these may be combined or separated for different goals.

5.2.2 | Methods

Methods of the study are usually known at the outset and can usually be pre-registered in the typical format. However, it will be important to update this document regularly or to post addenda with timestamped entries indicating changes to methods, the rationale for the change, and the likely consequences for analysis. For instance, on the OSF, you can upload revised documents with the same filename as the original to over-write the previous document, and the OSF keeps a version history of each change with timestamps. Alternatively, you can upload the addenda in a separate document with a different filename so that others know that they are expected to read both the original document and the addenda. It can be helpful to share manuals, protocols, consent forms, and stimuli. Protocols of proprietary products may not be able to be shared, but researchers are encouraged to share whatever they can legally and ethically share.

5.2.3 | Hypotheses

Over the course of a longitudinal project, hypotheses will develop in response to new data or to new findings from other labs. Thus, we propose that hypotheses be registered independently of the index variables and the statistical models, and likely in multiple stages. Hypotheses should be registered as a standalone document that details the hypothesis and its empirical or theoretical basis, what was known from the study when the hypothesis was developed, what motivated it, and the logic of the indices or measures that are used to test it. Each hypothesis can be registered separately as they arise. Critically, this can avoid SHARKing, and can preserve a difference between confirmatory and exploratory analyses for use in family-wise error correction procedures later.

5.2.4 | The sample

While longitudinal researchers strive for a generalizable sample, the precise structure of the sample often cannot be known in advance. However, this affects hypotheses, index identification plans, and analysis. Thus, at the conclusion of recruiting and enrollment, we suggest a registration document that clearly describes the sample. This should include several key details. First, researchers should provide sufficient information to understand the actual sample tested, including age ranges, demographics, school grades, geographic setting, etc.

Second, it is important to document how participants were recruited. This can affect the population to be generalized, such as if recruiting methods inadvertently under-recruited particular groups.

Third, in a longitudinal study, missingness and attrition can be important issues and often do not occur completely at random. This could result in biased inferences and limits to generalizability. Analysis plans to account for missing data should be explicitly described. If missing data will be imputed, imputation strategies should be specified in advance. If missing data will be excluded, the rationale for excluding these data, and decisions on whether to exclude other measures for that same participant should be detailed. Ideally, all the papers deriving from a longitudinal study would use the same general strategy for handling missing data given a particular analytic approach (e.g., multiple imputations for mixed models, and full information likelihood for factor analysis). Thus, the registration document could lock in the parameters for these procedures, such as the number of iterations and the types of

variables included in the imputation models. The researcher can also specify the expected amount of missing data and plans to prevent missing data.

Fourth, it is important to describe the population to whom the findings would most likely generalize, the extent to which the findings may generalize beyond the study context, and potential constraints on the generalizability of findings (Barbot et al., 2020).

5.2.5 | Identification and computation of indices

Longitudinal studies often include multiple measures intended to assess the same construct and are combined into what we term an index. Factor analytic approaches depend on the data structure. Similarly, studies using complex measures like eye-tracking or EEG may present researchers with multiple opportunities for deciding which measure(s) to use to index the target construct (Box 2). This too may depend on the data. For example, if a non-linear curve fitting approach is used (c.f., McMurray et al., 2010), the data may not fit the predicted function. These situations can be challenging for traditional (complete) pre-registration of both methods and analysis, because key variables may not be definable until after the data are collected.

Thus, we propose a separate registration document for how measures are identified and developed. This should specify the nature of the raw data that contribute, a plan for the approach that will be used for data reduction (e.g., factor analysis, item response theory, or a curve fit), and a set of hypothetical decisions for deciding that a given measure is optimal. It should clearly state whether such decisions will be based on metrics that are internal to the data/index (e.g., fit of confirmatory factor analysis, reliability), whether they will be based on criterion-related validity to an external index, and/or whether they were decided a priori on the basis of prior work or theoretical concerns. This is important because it makes it clear to the reader the extent to which these decisions may be biased toward or against hypotheses. This can be defined separately for each measure. This can be quite simple (e.g., 'We will use a difference score because that is what is commonly used with the Flanker task')—the important thing is to state the index, the plan for evaluation, and the motivation clearly before analysis.

5.2.6 | Analysis

Finally, analyses can be pre-registered separately from these other documents. These should reference the other pre-registration documents, and indeed they will depend on them—analysis plans should be linked to the hypotheses and the indices. This way, analyses can be developed as the hypothesis unfolds. From our perspective, the important thing is not that these are registered in advance, but rather that they clearly and honestly state several key points.

First, it is important to report whether the analysis will test a confirmatory or exploratory question. By clearly linking to the hypothesis registration, this should be straightforward. Second, it is important to state at what point in the project the analysis was determined. Third, it is valuable to state what was known about the data when the model was decided. Fourth, it is helpful to provide a priori power analyses for tests of the focal hypotheses. Then, the researcher can indicate what their approach would be if they do not achieve the target power—for example, due to a smaller sample or to greater missingness than anticipated. For instance, a researcher may choose to modify or preclude an analysis if adequate power is not achieved.

5.2.7 | Getting started

Before the study begins, we encourage researchers to provide an initial registration of the focal methods and research questions, which may be vague. If possible, one should present the hypotheses, along with the intended

analysis plan, although this is not always necessary (see Box 3). This pre-registration is likely to be less comprehensive than the typical pre-registration for single-session cross-sectional studies, for which exact details can be planned.

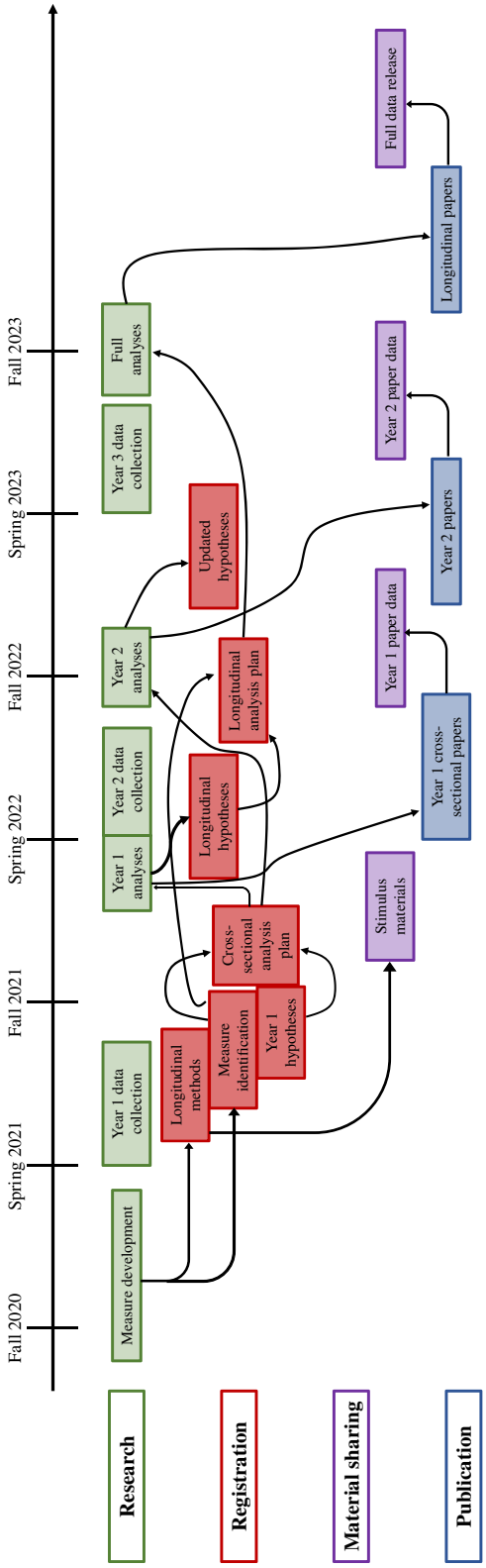
To this end, aspects of design and analysis should be described during pre-registration at an appropriate level of specificity to constrain researcher degrees of freedom. For example, although specific model structures (e.g., a growth curve versus a cross-lagged model) may not be possible to pre-specify, it may be possible to pre-specify more general principles such as the intended factors. It can also be helpful to register a framework for decision making including the considerations or decision tree that would lead one to take one approach versus another. Moreover, the researcher can specify how they would proceed if their measures do not demonstrate longitudinal factorial invariance, or how they would handle poorly fitting models. As is common in longitudinal studies, additional measures, research questions, and hypotheses may be added along the way, but it is helpful to provide the initial pre-registration with the core methods, aims, and hypotheses that motivated the study.

Next, as the project begins, researchers should begin developing more detailed hypotheses, index selection, and analysis plans. These should be registered before they are implemented. For instance, the researchers should decide on the index development plan before examining that aspect of the data. However, if indices are developed during implementation this can also be registered and noted. These registrations should be released and 'locked in' as they are finalized. Afterward, as additional decisions are made (e.g., measures are added or amended), questions or hypotheses are developed, and analysis plans are formalized, documents can be added or amended with a timestamp

BOX 3 Timeline of registration and data/materials sharing for two longitudinal studies

School readiness study timeline: The School Readiness Study began as a pilot project in 2018 with 62 children. In June 2020, the project received NIH funding to recruit additional children and to continue the project (current $N = 108$). From March 2020 to April 2021, lab visits were suspended due to the COVID-19 pandemic. In January 2021, we pre-registered our focal hypotheses (https://osf.io/gpn5y) and measures (https://osf.io/vg9jy) on the Open Science Framework (OSF). Later, we adapted our procedures to mitigate risk so we could safely collect data during the pandemic and adjusted our measures, accordingly. We also added measures to capture the effects of the pandemic on children. We thus made modifications to our pre-registered methods before data collection resumed in May 2021. We are currently creating a Data Dictionary of composite variables in the study (https://osf.io/e62uq) to help outside researchers—as well as those on the team—understand the meaning of our variables. For various papers we are writing, we submitted secondary pre-registrations of hypotheses, methods, and analysis plans that were separate from the focal pre-registration for the project. When we submit papers for review, we will share the data (i.e., variables) used for that paper, a data dictionary for those variables, the analysis code, and a computational notebook.

Growing words timeline: The *Growing Words* Project ($N = 242$) was interrupted by the COVID-19 pandemic in the very early stages of data collection. While data collection was shut down, our team adjusted our protocols to move some measures to online testing, refine and/or shorten measures, and identify better candidate measurement tools. This resulted in a less traditional registration strategy, in which many of our initial hypotheses and methods were not registered until during or after the first year of data collection—but before analyses. Below is the registration strategy thus far, as well as the intended strategy moving forward. This diagram includes various components of the research process, including planned times for registration of methods, hypotheses, and analysis plans; for release of data and materials; and for publication of findings. The lines connecting components signify how completion of one component informs another.



associated with each version. In this framework, new registrations will follow as the study unfolds. For instance, follow-up registrations might include specific indices that will be examined or secondary hypotheses to be tested. These follow-up registrations can reach a level of specificity that may not have been feasible at the initial pre-registration. There is still value in registering hypotheses and analysis plans before they are conducted, even if data collection has begun because registration at this stage constrains analysis decisions and reduces the risk of *p*-hacking. It is crucial at this stage that researchers report the rationale for choices. Because some of these decisions may be made after data are available, a concrete record of why a decision was made minimizes researcher degrees of freedom, much like true pre-registration.

There are multiple outlets for registration, depending on the domain, including the OSF, PROSPERO (for systematic reviews), and AsPredicted. Currently, there is not a way to implement this easily as part of a formal pre-registration template such as the OSF's (e.g., which locks in document versions, assigns DOIs to registrations, etc.). Moreover, if this approach becomes in widespread use, it may be useful to develop tools that can not only do version control and content locking but can also provide a visualization of the timeline of pre-registration and automatically maintain hyperlinks to related documents. However, this can easily be implemented less formally as a collection of documents posted to a public repository like the OSF. We provide examples of registration of measures and hypotheses on the OSF for the longitudinal *School Readiness Study* (https://osf.io/jzxb8), and for the longitudinal *Growing Words* project (https://osf.io/vzb2k), which illustrate two approaches to implementing this vision. Both projects are underway, and in both, registration is dynamic and unfolding.

5.3 | Data/materials sharing

Data sharing in a longitudinal project requires special consideration. The American Psychological Association Data Sharing Workgroup (2015) suggests a data-sharing embargo window "commensurate with the research team's investment of effort in study conceptualization and implementation, as well as with the time required for the research team to conduct its own analyses of the data". Thus, we propose a reasonable embargo for longitudinal designs to avoid disincentivizing them.

Before making data sharing plans, we encourage researchers to get permission from the IRB and participants to share de-identified data, or at least non-sensitive data. Researchers may consider re-randomizing participant identification numbers in the shared data file across studies so participants cannot be linked across papers (see Walsh et al., 2018 for additional procedures for de-identifying sensitive data in clinical psychology). When de-identified data cannot be traced back to the individual, those data can be shared as a routine matter. However, if the data include extensive personal information like income, zip code, school attended, etc., this may challenge anonymization. In this case, it may be necessary to ask participants to share their data as part of the consent process; however, keep in mind the potential for sampling bias (Eisenberg, 2015): people from vulnerable and historically disadvantaged groups (such as women, African Americans, Latino/a Americans, gender and sexual minorities, and people from lower socioeconomic status backgrounds) may be less comfortable sharing their data. These and other historically disadvantaged groups have been systematically under-represented in research, leading to knowledge and interventions that disproportionately benefit more advantaged groups, and ultimately, health disparities (Kwiatkowski et al., 2013). Thus, it is critical to avoid under-sampling members of these groups in the name of open science. Video of participants can be shared using permission separate from the consent process, minimizing sample bias (see example language: https://databrary.org/support/irb/release-template.html).

We recommend sharing data when papers are submitted (Morey et al., 2016) or published. This should include—at a minimum—the de-identified data and variables used for that particular paper and linked to that paper (O. Klein, Hardwicke, et al., 2018) so that researchers can verify the results in the paper (Gilmore & Qian, 2021). This differs from the typical model of sharing *all* data but helps minimize the barrier of the vast effort it takes to curate all the data and helps researchers feel in control of their data to avoid being scooped. To avoid cherry-picking or

accusations thereof, this is most effective in the context of clear registrations of the nature of the entire dataset, and explanations of why the subset of data shared is appropriate. Although this does not provide the complete repository that is useful for secondary research (though see below), it does provide opportunities for others to replicate the published analyses, and to explore secondary issues raised by those papers.

In addition to the data, it is important to share other related materials as well. Metadata—that is, documentation about data such as codebooks and data dictionaries (Buchanan et al., 2021)—should be included so that the data can be used most effectively. Some repositories such as Databrary allow sharing protocol materials and an overview of the data (e.g., the number of files) to indicate what data will be shared, while data are embargoed. Given the burdens of organizing data and creating metadata to be useful for others (i.e., curating data), we recommend active curation (Soska et al., 2021): curating data as they are collected rather than waiting until they are posted. We also encourage researchers to share analysis scripts with code comments to help others reproduce analyses and findings. Computational notebooks, such as R Markdown (Xie et al., 2021) and Jupyter (Kluyver et al., 2016) notebooks, can weave together text, analysis code, and results inline.[2] Computational notebooks, along with interactive tools such as Shiny R apps (Lafit et al., 2021), provide a contemporary extension of scientific papers (Somers, 2018).

Particularly in this modular registration model, it is likely that registration documents and analysis plans may change. Consequently, it may be helpful to implement version control.[3] However, even without version control tools—which may be difficult for some to learn and maintain—it is important that the researcher maintains transparency about the nature, timeline, and rationale of changes, regardless of the platform used.

Data can be shared as supplemental material linked to the paper, in a journal that specializes in publishing data (e.g., *Scientific Data*), or in a data repository (Gilmore & Qian, 2021). A list of example data repositories is in Table 2. Open-access repositories are freely available to anyone and thus may be preferred for de-identified data. Restricted-access repositories use secure mechanisms for sharing data and have access restrictions, such as researcher training, or IRB approval. Thus, restricted-access repositories may better handle sensitive and identifiable data (Gilmore et al., 2020). For a step-by-step guide to pre-registration and data sharing, see Krypotos et al. (2019).

After the full longitudinal data collection is complete, researchers may consider sharing the whole data set, possibly after some embargo period. This can be decided in advance and included in the initial pre-registration document. Alternatively, data can be shared iteratively as the study proceeds, which has been called born-open data (Rouder, 2016). For example, data could be automatically uploaded daily to a public repository during data collection for increased transparency and to serve as an off-site backup (O. Klein, Hardwicke, et al., 2018).

5.4 | Reporting

Our recommendations around open science practices are premised on the fact that any given hypothesis, index, or analysis is almost always a subset of the full scope of a longitudinal project. This fact also applies to both publications and other forms of reporting, and it is worth briefly discussing the unique ways in which published longitudinal studies can support rigor, as well as ways in which this model can be applied to other types of work.

5.4.1 | Publications

Space limitations are often restrictive in journals, so researchers should provide the necessary information in supplementary materials or public repositories such as the OSF. This is another benefit of our modular registration model—even if the statistical model will not be pre-registered, the registration of the methods, index development plans, and so forth can provide this critical supplementary information. Text-based descriptions of procedures and stimuli may not be sufficient for replicability; original stimuli as well as videos of lab setup, procedures (e.g., computer-based tasks), and participant behaviors can help (Gilmore & Adolph, 2017). Videos not only enhance replicability they can

TABLE 2 Examples of data repositories

Data repository	Type of data submissions allowed					
	Open access	Restricted access	Private access or data embargo	Maximum embargo period	For more information	URL
Databrary	x	x	x	unlimited	Simon et al. (2015)	https://nyu.databrary.org
The Dataverse Project	x	x	x	unlimited	King (2007)	https://dataverse.org
Dryad	x		x	10 years	White et al. (2008)	https://datadryad.org
FigShare	x		x	unlimited		https://figshare.com
Inter-University Consortium for Political and Social Research (ICPSR)	x	x	x	3 years	Swanberg (2017)	https://www.icpsr.umich.edu/web/pages/deposit/index.html
Leibniz Institute for the Social Sciences (GESIS)	x	x	x	2 years	Schumann and Mauer (2013)	https://www.gesis.org/en/services/archiving-and-sharing/sharing-data
Open Science Framework (OSF)	x		x	unlimited (4 years for registrations)	Foster and Deardorff (2017)	https://osf.io
Qualitative Data Repository (QDR)		x	x	typically 1–3 years	Karcher et al. (2016)	https://qdr.syr.edu
TalkBank	x	x	x	negotiable (typically 1–2 years)	MacWhinney (2007)	https://www.talkbank.org
Zenodo	x	x	x	unlimited		https://zenodo.org

Note: Many research institutions also have data repositories. In an open-access (aka public-access) repository, the data are available for download by the public, without restrictions. In a restricted-access repository, the data are available for download by people who meet the proper restrictions; as one example: researchers from authorized institutions who sign a data-use agreement. In a private-access (aka closed-access) repository, the data are available for download by only the investigators who posted the data, or by the specific people to whom they give access. An embargo is the period during which the uploaded data are not available for download by others. A private-access repository can function as a data embargo, in which the embargo ends when the researcher changes the repository access from private access to open or restricted access.

make it less likely that abundant data from participants go to waste upon publication (Adolph et al., 2012; Gilmore & Adolph, 2017), and Databrary can enable this.

Additional information is also important to report for transparency. Publications should describe which ages the broader longitudinal study spanned, which ages were examined in the paper, and why those ages were selected. It is helpful to report all predictors (e.g., independent variables) and outcomes (dependent variables) that are relevant to the paper—not just those that support hypotheses. It is also important to describe the degree of missingness and attrition, what led to missing data, and to provide tests that examine whether missingness in variables of interest is systematic (Nicholson et al., 2017). Publications should also explicitly point to pre-registration documents detailing the study, especially for aspects of the study not covered by that publication.

The broader goal of open science is to improve rigor. Although sharing data and materials is important for this, publications should consider the full range of causes of errors in scientific inference. Some inference errors could, in part, reflect measurement error (Loken & Gelman, 2017). Reliability and validity depend on the context and sample. It is insufficient for papers to state that the measures are reliable and valid; instead, multiple aspects of reliability and validity should be described, especially with respect to the sample and population of interest, and researchers should consider ways to improve the reliability and validity. For instance, latent factors could be assessed using multi-trait multi-method assessments (Patrick et al., 2013). This dovetails neatly with modular approaches to registering index development plans. Other replication failures could, in part, reflect low power. Thus, it is important to report effect sizes and not just statistical significance (Klapwijk et al., 2021).

5.4.2 | Other types of work

Critically, this modular pre-registration is ideal for a range of possible longitudinal studies. For example, accelerated longitudinal designs may track children from multiple starting ages at the same time. This permits cross-sectional comparisons in Year 1, whose hypotheses and statistical approaches can be registered separately from longitudinal analyses in later years. Other longitudinal studies may piggy-back on contract work whose goals are primarily to address an empirical question needed by a funder. Here, the investigators could register the methods and key indices initially, and perhaps later register more hypothesis-driven analyses as they arise.

Modular registration is even appropriate for secondary data analysis such as datasets described earlier (ABCD, Add Health, ECLS, NLS, NICHD SECCYD), because it separates hypotheses and models from the methods. Analysis of preexisting data sets can be registered (Weston et al., 2019), and some journals accept secondary registered reports that allow registration after data have been collected but before data analysis (Pfeifer & Weston, 2020). The modular approach proposed here is ideally suited to this enterprise as it allows the registration of hypotheses or analyses to be independent of other things; indeed, in some ways what we are advocating is akin to what people are already doing with secondary datasets. There may be additional challenges that need to be registered in contexts like this, such as procedures for extracting and filtering the relevant data, but this can fit cleanly into our framework.

5.5 | What we would like to see

Publication and funding incentives must better align with scientific rigor. We are encouraged that some journals accept replication studies (ReScienceX; Roesch & Rougier, 2020), null findings (e.g., *Journal of Articles in Support of the Null Hypothesis*), and registered reports (Nosek & Lakens, 2014). Nevertheless, more—or, ideally, all—journals should allow and encourage these things. We encourage research institutions to value open science practices in hiring and promotion decisions.

In the future, it would be good to see journals and funding agencies value the important role of exploratory work in addition to confirmatory work. It can slow science to subject exploratory work to the same time-consuming

processes that we expect of confirmatory work. However, in this context, it is important to see more explicit, thorough, and genuine discussion in papers of the extent to which research questions are exploratory or confirmatory (Davis-Kean & Ellis, 2019). As we have illustrated here, longitudinal work blurs the line between exploratory and confirmatory work, and the modular and time-extended approach to registration can serve an important goal of clarifying the nature of individual research questions and analyses in a longitudinal study or any other large individual differences study. In the context of truly exploratory work or analyses of existing data, our modular approach may have value in allowing researchers to register some things independently of others without being overly constraining.

These are expectations that should affect all areas of science. However, as longitudinal studies illustrate, there is not a one-size-fits-all approach to pre-registration and data sharing. We encourage journal editors, faculty review committees, and students to embrace a variety of models of data-sharing and registration, including the possibility that sharing and/or registration may be unwarranted or unnecessary in some circumstances, such as for purely exploratory questions, for low-stakes or early-stage studies that are not testing large theories, etc. Critically, the goal should not be to dogmatically adhere to a particular set of practices but to be open about what was done, what was known at the time, and what factors led to the scientific decisions.

Finally, we point out that the technical demands of open science are daunting—even just to follow the guidelines in this paper, it may appear that a researcher will need to retool their lab to learn an array of new acronyms: to master R markdown and GitHub, figure out a system for video tagging, IRB procedures for data sharing, and several repositories for embargoing. It is daunting. While these technical tools are valuable, the important thing is, to be honest, clear, and open. Simply posting PDFs of registration documents to a public repository, a journal article's supplement, or even your lab website is better than not sharing at all.

6 | OPEN SCIENCE IS NOT ENOUGH

Although we present these guidelines with aims to advance the open science movement and to increase rigor, open science by itself is insufficient for this goal. Much discussion has focused on questionable research practices and lack of transparency, but that is likely not the only contributor to the replication crisis. What will make developmental science more replicable is better science, including a better understanding of constructs and more valid assessments of those constructs. Unreliability of measures could be a key contributor to the replication crisis (Loken & Gelman, 2017), because the reliability of a measure is the upper limit of its validity. Many standard cognitive measures have poor test–retest reliability (Enkavi et al., 2019; Hedge et al., 2018). The reliability of measures is especially crucial for longitudinal studies because the reliability of change scores (or difference scores) is lower than the reliability of the measure at each timepoint (Revelle & Condon, 2019). Thus, what may appear to be a change in people's scores from T1 to T2 may in many cases reflect measurement error rather than changes in the person's level on the construct. In addition, there is often flexibility in mapping hypotheses to measures or to differences among conditions. A 'weak logical link between theory and their empirical tests' can lead to an inflation of Type I error or to weak tests that fail to detect a true effect (Oberauer & Lewandowsky, 2019). This often derives from a lack of understanding of the *derivation chain* (Meehl, 1990; Scheel et al., 2021)—the chain of the assumption that links an underlying theory to observed behavior. Thus, we caution readers to select measures that are highly reliable and tightly coupled to theory.

7 | CONCLUSION

There is no one-size-fits-all open science approach across scientific methods, and it has been particularly difficult to find one for longitudinal work. We propose a serial and modular approach to registration that includes an initial pre-registration of focal methods and hypotheses, along with subsequent pre- or co-registered questions and hypotheses

associated with specific papers—including exploratory work. We also encourage researchers to share their research materials and relevant data with associated papers, but released gradually and tied to publications and to specific goals. At the same time, we encourage funding agencies, research institutions, and stakeholders to think carefully about requirements regarding the timing of data sharing to avoid disincentivizing researchers from conducting important longitudinal studies.

ACKNOWLEDGEMENTS

Measures and hypotheses for the School Readiness Study were pre-registered: https://osf.io/jzxb8. Hypotheses for the Growing Words project were pre-registered: https://osf.io/vzb2k. The School Readiness Study (Study #: 201708761) and Growing Words project (Study #: 201809789) were approved by the University of Iowa Institutional Review Board. The School Readiness Study (Petersen and McMurray) was funded by Grants HD098235 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) and UL1TR002537 from the National Center for Advancing Translational Sciences (NCATS). The Growing Words project (McMurray, Apfelbaum, and Petersen) was funded by Grant DC008089 from the National Institute on Deafness and Other Communication Disorders (NIDCD).

CONFLICT OF INTEREST

We have no conflicts of interest to disclose.

AUTHOR CONTRIBUTIONS

Isaac T. Petersen: Conceptualization; funding acquisition; investigation; methodology; writing – original draft; writing – review and editing. **Keith S. Apfelbaum:** Conceptualization; funding acquisition; investigation; methodology; visualization; writing – review and editing. **Bob McMurray:** Conceptualization; funding acquisition; investigation; methodology; writing – review and editing.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/icd.2315.

ORCID

Isaac T. Petersen https://orcid.org/0000-0003-3072-6673
Keith S. Apfelbaum https://orcid.org/0000-0001-6955-6574
Bob McMurray https://orcid.org/0000-0002-6532-284X

ENDNOTES

[1] Registered reports have multiple benefits. First, after a registered report is approved, the research is likely to be published regardless of the findings, which combats the file-drawer problem. Second, registered reports ensure that reviewers agree that the proposed experiment is the right way to test the proposed hypotheses. However, given the scale and timeline of longitudinal studies, these are likely to be rarely used.

[2] For an example of a computational notebook associated with one of our papers (under review), see the R Markdown file (https://osf.io/sfyq7/?view_only=6aeae4e3f3f844fa94956c695bacac2d) and the associated html notebook: https://osf.io/xp2rs/?view_only=6aeae4e3f3f844fa94956c695bacac2d.

[3] Version control using software such as GitHub (Gilroy & Kaplan, 2019) can be used to share protocols (e.g., https://www.play-project.org) and analysis code, and track changes. The use of analysis syntax or code rather than point-and-click graphical user interfaces leads to better reproducibility (Gilmore et al., 2020). Using plain text documents yields the greatest benefits of versioning using GitHub. Examples of plain text documents include Markdown text (.md), analysis scripts (e.g., R: .r or .rmd; SPSS: .sps; SAS: .sas; STATA: .do; MATLAB: .mat; Python: .py; Mplus: .inp), and other text documents (e.g., .txt). A key benefit of plain text documents rather than binary files such as Word or PDF documents is that GitHub can track specific changes to plain text documents with timestamps, for a complete version history.

REFERENCES

Adolph, K. E., Gilmore, R. O., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward open behavioral science. *Psychological Inquiry*, *23*(3), 244–247. https://doi.org/10.1080/1047840X.2012.705133

APA Data Sharing Working Group. (2015). *Data sharing: Principles and considerations for policy development*. https://www.apa.org/science/leadership/bsa/data-sharing-report.pdf

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Barbot, B., Hein, S., Trentacosta, C., Beckmann, J. F., Bick, J., Crocetti, E., Liu, Y., Rao, S. F., Liew, J., Overbeek, G., Ponguta, L. A., Scheithauer, H., Super, C., Arnett, J., Bukowski, W., Cook, T. D., Côté, J., Eccles, J. S., Eid, M., … van IJzendoorn, M. H. (2020). Manifesto for new directions in developmental science. *New Directions for Child and Adolescent Development*, *2020*(172), 135–149. https://doi.org/10.1002/cad.20359

Bell, M. A., & Cuevas, K. (2012). Using EEG to study cognitive development: Issues and practices. *Journal of Cognition and Development*, *13*(3), 281–294. https://doi.org/10.1080/15248372.2012.691143

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—When and how? *Journal of Clinical Epidemiology*, *54*(4), 343–349. https://doi.org/10.1016/S0895-4356(00)00314-0

Benning, S. D., Bachrach, R. L., Smith, E. A., Freeman, A. J., & Wright, A. G. C. (2019). The registration continuum in clinical science: A guide toward transparent practices. *Journal of Abnormal Psychology*, *128*(6), 528–540. https://doi.org/10.1037/abn0000451

Berkman, B. E., Howard, D., & Wendler, D. (2018). Reconsidering the need for reconsent at 18. *Pediatrics*, *142*(2), 1–5. https://doi.org/10.1542/peds.2017-1202

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*, 435.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Brooker, R. J., Bates, J. E., Buss, K. A., Canen, M. J., Dennis-Tiwary, T. A., Gatze-Kopp, L. M., Hoyniak, C. P., Klein, D. N., Kujawa, A., Lahat, A., Lamm, C., Moser, J. S., Petersen, I. T., Tang, A., Woltering, S., & Schmidt, L. A. (2020). Conducting event-related potential (ERP) research with young children: A review of components, special considerations and recommendations for research on cognition and emotion. *Journal of Psychophysiology*, *34*(3), 137–158. https://doi.org/10.1027/0269-8803/a000243

Brown, S. D., Furrow, D., Hill, D. F., Gable, J. C., Porter, L. P., & Jacobs, W. J. (2014). A duty to describe: Better the devil you know than the devil you don't. *Perspectives on Psychological Science*, *9*(6), 626–640. https://doi.org/10.1177/1745691614551749

Buchanan, E. M., Crain, S. E., Cunningham, A. L., Johnson, H. R., Stash, H., Papadatou-Pastou, M., Isager, P. M., Carlsson, R., & Aczel, B. (2021). Getting started creating data dictionaries: How to create a shareable data set. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920928007. https://doi.org/10.1177/2515245920928007

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637–644. https://doi.org/10.1038/s41562-018-0399-z

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., … Dale, A. M. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, *32*, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001

Chase-Lansdale, P. L., Mott, F. L., Brooks-Gunn, J., & Phillips, D. A. (1991). Children of the National Longitudinal Survey of Youth: A unique research opportunity. *Developmental Psychology*, *27*(6), 918–931. https://doi.org/10.1037/0012-1649.27.6.918

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), e13437. https://doi.org/10.1111/psyp.13437

Craig, R., Cox, A., Tourish, D., & Thorpe, A. (2020). Using retracted journal articles in psychology to understand research misconduct in the social sciences: What is to be done? *Research Policy*, *49*(4), 103930. https://doi.org/10.1016/j.respol.2020.103930

Davis-Kean, P. E., & Ellis, A. (2019). An overview of issues in infant and developmental research for the creation of robust and replicable science. *Infant Behavior and Development*, *57*, 101339. https://doi.org/10.1016/j.infbeh.2019.101339

de Groot, A. D. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. *Acta Psychologica*, *148*, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001

Eisenberg, N. (2015). Thoughts on the future of data sharing. *APS Observer*, *28*(5). https://www.psychologicalscience.org/observer/thoughts-on-the-future-of-data-sharing

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*, *10*, e67995. https://doi.org/10.7554/eLife.67995

Fellman, D., Jylkkä, J., Waris, O., Soveri, A., Ritakallio, L., Haga, S., Salmi, J., Nyman, T. J., & Laine, M. (2020). The role of strategy use in working memory training outcomes. *Journal of Memory and Language*, *110*, 104064. https://doi.org/10.1016/j.jml.2019.104064

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, *105*(2), 4. https://doi.org/10.5195/jmla.2017.88

Francis, G., & Thunell, E. (2021). Reversing Bonferroni. *Psychonomic Bulletin & Review*, *28*(3), 788–794. https://doi.org/10.3758/s13423-020-01855-z

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Gelman, A. (2016). Why does the replication crisis seem worse in psychology? *Slate*.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Department of Statistics, Columbia University.

Gennetian, L. A., Tamis-LeMonda, C. S., & Frank, M. C. (2020). Advancing transparency and openness in child development research: Opportunities. *Child Development Perspectives*, *14*(1), 3–8. https://doi.org/10.1111/cdep.12356

Gilmore, R. O. (2016). From big data to deep insight in developmental science. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(2), 112–126. https://doi.org/10.1002/wcs.1379

Gilmore, R. O., & Adolph, K. E. (2017). Video can make behavioural science more reproducible. *Nature Human Behaviour*, *1*, 0128. https://doi.org/10.1038/s41562-017-0128

Gilmore, R. O., Cole, P. M., Verma, S., van Aken, M. A. G., & Worthman, C. M. (2020). Advancing scientific integrity, transparency, and openness in child development research: Challenges and possible solutions. *Child Development Perspectives*, *14*(1), 9–14. https://doi.org/10.1111/cdep.12360

Gilmore, R. O., & Qian, Y. (2021). An open developmental science will be more rigorous, robust, and impactful. *Infant and Child Development*, *31*, e2254. https://doi.org/10.1002/icd.2254

Gilmore, R. O., Xu, M., & Adolph, K. E. (2021). Data sharing. In S. Panecker & B. Stanley (Eds.), *Handbook of research ethics in psychological science*. American Psychological Association.

Gilroy, S. P., & Kaplan, B. A. (2019). Furthering open science in behavior analysis: An introduction and tutorial for using GitHub in research. *Perspectives on Behavior Science*, *42*(3), 565–581. https://doi.org/10.1007/s40614-019-00202-5

Hassan, M. A., Malik, A. S., Fofi, D., Saad, N., Karasfi, B., Ali, Y. S., & Meriaudeau, F. (2017). Heart rate estimation using facial video: A review. *Biomedical Signal Processing and Control*, *38*, 346–360. https://doi.org/10.1016/j.bspc.2017.07.004

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hepach, R., & Westermann, G. (2016). Pupillometry in infancy research. *Journal of Cognition and Development*, *17*(3), 359–377. https://doi.org/10.1080/15248372.2015.1135801

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management*, *43*(1), 5–18. https://doi.org/10.1177/0149206316679487

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Karcher, S., Kirilova, D., & Weber, N. (2016). Beyond the matrix: Repository services for qualitative data. *IFLA Journal*, *42*(4), 292–302. https://doi.org/10.1177/0340035216672870

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods & Research*, *36*(2), 173–199. https://doi.org/10.1177/0049124107306660

Kirkman, B. L., & Chen, G. (2011). Maximizing your data or data slicing? Recommendations for managing multiple submissions from the same dataset. *Management and Organization Review*, *7*(3), 433–446. https://doi.org/10.1111/j.1740-8784.2011.00228.x

Klapwijk, E. T., van den Bos, W., Tamnes, C. K., Raschle, N. M., & Mills, K. L. (2021). Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Developmental Cognitive Neuroscience*, *47*, 1–19. https://doi.org/10.1016/j.dcn.2020.100902

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), 20. https://doi.org/10.1525/collabra.158

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Jupyter Development Team. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. *20th International Conference on Electronic Publishing (01/01/16)*. https://eprints.soton.ac.uk/403913/.

Krypotos, A.-M., Klugkist, I., Mertens, G., & Engelhard, I. M. (2019). A step-by-step guide on preregistration and effective data sharing for psychopathology research. *Journal of Abnormal Psychology*, *128*(6), 517–527. https://doi.org/10.1037/abn0000424

Kwiatkowski, K., Coe, K., Bailar, J. C., & Swanson, G. M. (2013). Inclusion of minorities and women in cancer clinical trials, a decade later: Have we improved? *Cancer*, *119*(16), 2956–2963. https://doi.org/10.1002/cncr.28168

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly Shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920978738. https://doi.org/10.1177/2515245920978738

Law, F., Mahr, T., Schneeberg, A., & Edwards, J. A. N. (2017). Vocabulary size and auditory word recognition in preschool children. *Applied PsychoLinguistics*, *38*(1), 89–125. https://doi.org/10.1017/S0142716416000126

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

MacWhinney, B. (2007). The Talkbank project. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora Synchronic databases* (Vol. 1, pp. 163–180). Palgrave Macmillan UK. https://doi.org/10.1057/9780230223936_7

McMurray, B., Apfelbaum, K. S., & Tomblin, J. B. (in press). The slow development of real-time processing: Spoken word recognition as a crucible for new about thinking about language acquisition and disorders. *Current Directions in Psychological Science*.

McMurray, B., Ellis, T. P., & Apfelbaum, K. S. (2019). How do you deal with uncertainty? Cochlear implant users differ in the dynamics of lexical processing of noncanonical inputs. *Ear and Hearing*, *40*(4), 961–980. https://doi.org/10.1097/aud.0000000000000681

McMurray, B., Samelson, V. M., Lee, S. H., & Bruce Tomblin, J. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology*, *60*(1), 1–39. https://doi.org/10.1016/j.cogpsych.2009.06.003

Meehl, P. E. (1990). Why summaries of research on psychological theories are often unintrepretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. https://doi.org/10.1177/2515245917747656

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., Lewandowsky, S., Morey, C. C., Newman, D. P., Schönbrodt, F. D., Vanpaemel, W., Wagenmakers, E.-J., & Zwaan, R. A. (2016). The peer reviewers' openness initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547. https://doi.org/10.1098/rsos.150547

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. https://doi.org/10.1038/s41562-016-0021

NICHD Early Child Care Research Network. (2005). *Child care and child development: Results from the NICHD Study of Early Child Care and Youth Development*. Guilford Press.

Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development*, *41*(1), 143–153. https://doi.org/10.1177/0165025415618275

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Patrick, C. J., Venables, N. C., Yancey, J. R., Hicks, B. M., Nelson, L. D., & Kramer, M. D. (2013). A construct-network approach to bridging diagnostic and physiological domains: Application to assessment of externalizing psychopathology. *Journal of Abnormal Psychology*, *122*(3), 902–916. https://doi.org/10.1037/a0032807

Petersen, I. T., Choe, D. E., & LeBeau, B. (2020). Studying a moving target in development: The challenge and opportunity of heterotypic continuity. *Developmental Review*, *58*, 100935. https://doi.org/10.1016/j.dr.2020.100935

Pfeifer, J. H., & Weston, S. J. (2020). Developmental cognitive neuroscience initiatives for advancements in methodological approaches: Registered Reports and Next-Generation Tools. *Developmental Cognitive Neuroscience*, *44*, 100755. https://doi.org/10.1016/j.dcn.2020.100755

Resnick, B. (2018). *The Stanford Prison Experiment was massively influential. We just learned it was a fraud* [Vox]. Twitter.

Resnick, B. (2021). *The replication crisis devastated psychology. This group is looking to rebuild it* [Vox]. Twitter.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., Ireland, M., Bearinger, L. H., & Udry, J. R. (1997). Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *JAMA*, *278*(10), 823–832. https://doi.org/10.1001/jama.1997.03550100049038

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Rigler, H., Farris-Trimble, A., Greiner, L., Walker, J., Tomblin, J. B., & McMurray, B. (2015). The slow developmental time course of real-time spoken word recognition. *Developmental Psychology*, *51*(12), 1690–1703. https://doi.org/10.1037/dev0000044

Roesch, E., & Rougier, N. P. (2020). New journal for reproduction and replication results. *Nature*, *581*(7806), 30. https://doi.org/10.1038/d41586-020-01328-2

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science*, *5*(3), 127–134. https://doi.org/10.1111/j.1467-9280.1994.tb00646.x

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior Research Methods*, *48*(3), 1062–1069. https://doi.org/10.3758/s13428-015-0630-z

Rubin, M. (2017). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology*, *21*(3), 269–275. https://doi.org/10.1037/gpr0000123

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schumann, N., & Mauer, R. (2013). The GESIS data archive for the social sciences: A widely recognised data archive on its way. *International Journal of Digital Curation*, *8*(2), 215–222.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simon, D. A., Gordon, A. S., Steiger, L., & Gilmore, R. O. (2015). Databrary: Enabling sharing and reuse of research video. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, Knoxville, Tennessee, USA*. https://doi.org/10.1145/2756406.2756951

Somers, J. (2018). The scientific paper is obsolete. *The Atlantic*, 4. https://www.theatlantic.com/science/archive/2018/04/the-scientific-paper-is-obsolete/556676/

Soska, K. C., Xu, M., Gonzalez, S. L., Herzberg, O., Tamis-LeMonda, C. S., Gilmore, R. O., & Adolph, K. E. (2021). (Hyper)active data curation: A video case study from behavioral science. *Journal of eScience Librarianship*, *10*(3), e1208. https://doi.org/10.7191/jeslib.2021.1208

Strand, J. (2020). *Scientists make mistakes. I made a big one*. https://elemental.medium.com/when-science-needs-self-correcting-a130eacb4235

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7*(6), 670–688. https://doi.org/10.1177/1745691612460687

Swanberg, S. M. (2017). Inter-university consortium for political and social research (ICPSR). *Journal of the Medical Library Association*, *105*(1), 2. https://doi.org/10.5195/jmla.2017.120

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology*, *15*(1), 579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710

Tackett, J. L., Brandes, C. M., & Reardon, K. W. (2019). Leveraging the Open Science framework in clinical psychological assessment research. *Psychological Assessment*, *31*(12), 1386–1394. https://doi.org/10.1037/pas0000583

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*(5), 742–756. https://doi.org/10.1177/1745691617690042

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. https://doi.org/10.1038/s41597-021-00981-0

Thompson, W. H., Wright, J., Bissett, P. G., & Poldrack, R. A. (2020). Dataset decay and the problem of sequential analyses on open datasets. *eLife*, *9*, e53498. https://doi.org/10.7554/eLife.53498

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K): Combined user's manual for the ECLS-K eighth-grade and K-8 full sample data files and electronic codebooks. NCES 2009-004. National Center for Education Statistics.

Walsh, C. G., Xia, W., Li, M., Denny, J. C., Harris, P. A., & Malin, B. A. (2018). Enabling open-science initiatives in clinical psychology and psychiatry without sacrificing patients' privacy: Current practices and future challenges. *Advances in Methods and Practices in Psychological Science*, *1*(1), 104–114. https://doi.org/10.1177/2515245917749652

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214–227. https://doi.org/10.1177/2515245919848684

White, H., Carrier, S., Thompson, A., Greenberg, J., & Scherle, R. (2008). The Dryad data repository: A Singapore framework metadata architecture in a DSpace environment. *Dublin Core Conference*.

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. https://doi.org/10.1037/0003-066X.61.7.726

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing*, *22*, 2331216518800869. https://doi.org/10.1177/2331216518800869

Xie, Y., Allaire, J. J., & Grolemund, G. (2021). *R Markdown: The definitive guide*. Chapman and Hall/CRC.

How to cite this article: Petersen, I. T., Apfelbaum, K. S., & McMurray, B. (2024). Adapting open science and pre-registration to longitudinal research. *Infant and Child Development*, *33*(1), e2315. https://doi.org/10.1002/icd.2315