Received: 6 February 2023 | Revised: 8 May 2023 | Accepted: 27 May 2023

DOI: 10.1111/psyp.14365

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Managing EEG studies: How to prepare and what to do once data collection has begun

Megan A. Boudewyn[1] | **Molly A. Erickson**[2] | **Kurt Winsler**[3] | **John Daniel Ragland**[4] | **Andrew Yonelinas**[3] | **Michael Frank**[5] | **Steven M. Silverstein**[6] | **Jim Gold**[7] | **Angus W. MacDonald III**[8] | **Cameron S. Carter**[4] | **Deanna M. Barch**[9] | **Steven J. Luck**[3]

[1]Department of Psychology, University of California Santa Cruz, Santa Cruz, California, USA

[2]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, Illinois, USA

[3]Department of Psychology, University of California, Davis, California, USA

[4]Department of Psychiatry and Behavioral Sciences, University of California, Davis, California, USA

[5]Department of Cognitive, Linguistics and Psychological Sciences, Brown University, Providence, Rhode Island, USA

[6]Department of Psychiatry, Neuroscience and Opthamology, University of Rochester, Rochester Medical Center, Rochester, New York, USA

[7]Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland, USA

[8]Department of Psychology, University of Minnesota, Minneapolis, Minnesota, USA

[9]Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, Missouri, USA

Correspondence
Megan A. Boudewyn, Department of Psychology, University of California Santa Cruz, Santa Cruz, CA 95064, USA.
Email: mboudewyn@ucsc.edu

Funding information
National Institute of Mental Health, Grant/Award Number: R01 MH084821, R01 MH084826, R01 MH084828, R01 MH084840, R01 MH084861 and R01 MH087450

Abstract

In this paper, we provide guidance for the organization and implementation of EEG studies. This work was inspired by our experience conducting a large-scale, multi-site study, but many elements could be applied to any EEG project. Section 1 focuses on study activities that take place before data collection begins. Topics covered include: establishing and training study teams, considerations for task design and piloting, setting up equipment and software, development of formal protocol documents, and planning communication strategy with all study team members. Section 2 focuses on what to do once data collection has already begun. Topics covered include: (1) how to effectively monitor and maintain EEG data quality, (2) how to ensure consistent implementation of experimental protocols, and (3) how to develop rigorous preprocessing procedures that are feasible for use in a large-scale study. Links to resources are also provided, including sample protocols, sample equipment and software tracking forms, sample code, and tutorial videos (to access resources, please visit: https://osf.io/wdrj3/).

KEYWORDS
EEG methods, guidelines, large-scale, multisite, protocol, recommendations

M. A. Boudewyn and M. A. Erickson equal contributions to the manuscript.

1 | OVERVIEW

There is a lot to consider when setting up and managing an electroencephalography (EEG) study, beyond the scientific hypothesis being tested. This is particularly true of large-scale and/or multi-site studies, which are becoming increasingly common in the present climate of collaborative science and emphasis on rigor and reproducibility (see Pavlov et al., 2021 for another recent approach). Although there are many obvious advantages to such studies, they inevitably introduce a number of implementation challenges that increase considerably as more participants, experimental tasks, follow-up sessions, data collection sites, and personnel become involved. Alongside a greater opportunity for error, the *cost* of those errors is commensurately larger. Imagine, for instance, that you spend several years planning and conducting a large-scale EEG study, and when data collection is complete you discover that some of your planned analyses cannot be performed because there were errors in the task presentation script, or you discover that the EEG was so noisy that your key effects are not statistically significant despite the large *N*. Similarly, you may find that there are large numbers of artifacts and you must therefore exclude many of your participants from the final analysis, causing your statistical power and generalizability—primary strengths of large studies—to suffer. Alternatively, you may be conducting a smaller-scale study and have a limited amount of time to collect a given number of high-quality datasets (e.g., for a dissertation project). For both large-scale and smaller-scale projects, it is well worth the time and effort to set up robust data collection systems and monitoring procedures to minimize errors and rapidly detect and correct problems when they arise. The set of suggestions that are most useful for a given study will depend on the scale and characteristics of that study.

This paper is motivated by the experience we gained over the course of conducting a large-scale, multi-site EEG study with the Cognitive Neurocomputational Task Reliability And Clinical Applications for Serious Mental Illness (CNTRACS) Consortium. This consortium was convened to identify tasks and computational models that meet the measurement standards necessary for clinical research on cognitive impairment in serious mental illness (Barch et al., 2009; Cohen & Insel, 2008; Luck & Gold, 2008). EEG was collected across five sites from 260 research participants, each of whom completed six experimental tasks, yielding a total of 1560 datasets. Thus, these recommendations focus on preparing for and managing the collection and preprocessing of a large number of datasets, as well as the additional challenges that arise when multiple research teams are involved in data collection. It is worth noting, however, that nearly all of these guidelines are applicable to the organization of any EEG project, regardless of the number of sites or scope of the study. For example, many investigators may find these suggestions useful in building a set of study protocol blueprints to be applied to all studies conducted in their lab, even if each individual experiment may not be considered to be "large-scale." It is also worth noting that these recommendations are our own suggestions rather than official publication guidelines; there are of course other approaches to conducting high-quality EEG studies at any scale (Pavlov et al., 2021). There are also formal guidelines put forth by Society for Psychophysiological Research with respect to EEG analysis and results reporting (Keil et al., 2014, 2022; Picton et al., 2000).

In Section 1, we provide suggestions for establishing the research teams, optimizing the experimental design, setting up the equipment, and training the research staff on EEG data acquisition procedures. In Section 2, our goal is to provide guidance for monitoring data quality and handling a large number of data sets once the study is underway. In particular, we will describe strategies for maintaining consistent and reliable implementation of experimental protocols, monitoring data quality, and developing preprocessing methods that are well suited to a large-scale study. Links to resources are also provided, including sample protocols, sample equipment and software tracking forms, sample code, and tutorial videos (to access resources, please visit: https://osf.io/wdrj3/).

2 | SECTION 1: BEFORE DATA COLLECTION HAS BEGUN

This section covers issues that should be addressed prior to data collection. This includes issues related to team structure, experiment design, equipment and software considerations, training new team members, and communication strategies. A central theme is that your protocol should be designed to minimize the need for even minor deviations over time (or across sites in a multi-site study). In our experience, small exceptions to a protocol compound into major headaches as the scale of the study increases. For example, imagine that the EEG file naming convention is sporadically violated due to unclear instructions, or a subset of participants has an extra EEG recording file without clear documentation about why the extra file exists. Correction of these seemingly minor deviations can require days of effort to track down and account for during data analysis or may produce invalid results. It is more efficient to design and test every element of your study from start to finish before beginning data collection to minimize the number of rules and exceptions that must be hastily constructed on the fly.

2.1 | Before data collection: Establish your teams

Before your study begins, consider how many people will be needed to coordinate the EEG data collection and monitoring activities so that you can budget accordingly. For larger studies, we recommend assembling three teams: a data collection team, a data preprocessing team, and a supervisory team. Data collection team members may also serve on the data preprocessing team; however, these activities require different skills and might be performed at different stages of the study, so we separate them here for the sake of clarity. The responsibilities assigned to each of the three teams are summarized below and covered in more detail Section 2. In a smaller study, you may have a single team (or even a "team of one"), but it is still essential to think carefully about each of these tasks.

2.1.1 | Data collection team

In addition to their primary responsibility of collecting the EEG data, we recommend that all members of the data collection team participate in regular meetings with the EEG supervisory team to review data quality (see Section 2.2: *Conduct Quality Control Assessments*). This team is also responsible for ensuring that (1) all data has been properly backed up at the end of each session, (2) all remarkable events that occurred during the recording session have been documented, and (3) any experiment updates disseminated by the EEG supervisory team have been thoroughly tested.

2.1.2 | Data preprocessing team

After data collection is underway, we suggest training a team of research assistants to do basic EEG preprocessing. Imagine you have 800 datasets (200 participants with 4 tasks each) that take an average of 30 min per dataset to preprocess. This amounts to 20 weeks of effort for someone who spends 50% of their time on EEG preprocessing. Training a team of research assistants to do the preprocessing is a much smaller time commitment, by comparison. Irrespective of the size of the dataset, the data preprocessing team should plan to dedicate a significant amount of their project effort to performing the preprocessing steps (see Section 2.3: *Standardize the Preprocessing Pipeline* for more details).

2.1.3 | EEG supervisory team

Finally, we recommend assembling an EEG supervisory team for a variety of activities ranging from troubleshooting data collection problems to training the team of preprocessors. This team is responsible for the following:

1. For multi-site studies, site visits will be needed to set up equipment (or verify that existing equipment is set up identically across sites), train staff on the data collection protocol, and ensure that everything is working properly (see *Personnel Training*, below). In our experience, an in-person visit by an experienced researcher is ideal to ensure consistency across sites, which in turn saves time later. If this is not possible, workable alternatives may include a combination of remote meetings and video recordings.
2. Someone must be on call during recording sessions to troubleshoot urgent messages from members of the data collection team who are experiencing a recording issue that requires an immediate response (e.g., a broken ground electrode, missing event codes; see *Keeping Lines of Communication Open*). This avoids the need to cancel and reschedule sessions, which is expensive in both time and money.
3. Immediately after the study begins, and any time after there has been a change to the data collection procedures or the task script, an experienced researcher should perform a deep inspection of the first several datasets to ensure that all elements of the task are working as expected. This is especially important for multi-site EEG studies, in which there is ample room for error and miscommunication.
4. Although the goal is to catch and fix any errors before data collection begins, errors may nevertheless occur, or new problems may arise for a number of reasons. Someone will need to dedicate time to fixing such errors for future participants and then developing a method for repairing the data that have already been recorded (see Section 2.4: *Prepare for the Unexpected* for more details).
5. For studies with several principal investigators, you may wish to designate one individual to consolidate information about data quality, number of datasets lost due to artifacts, preliminary results, etc. for investigator meetings. This member of the EEG supervisory team may also provide feedback on data retention directly to the data collection team so that they can make adjustments as needed.
6. Someone will need to oversee implementation and testing of any experiment updates (which may occur if an operating system is updated mid-study, for example). Again, this is a task that becomes much more burdensome with an increasing number of data collection sites, and personnel costs should be budgeted accordingly (see Section 2.4: *Prepare for the Unexpected* for more details).

7. Finally, the EEG supervisory team will be responsible for conducting quality control meetings with the data collection team and training the data preprocessing team. During the early phase of data collection and preprocessing for the CNTRACS project, the EEG supervisory team allocated 1 h per week for each of these activities.

Though this does not constitute an exhaustive list of every issue that may arise, the overarching message should be clear: for large-scale and/or multi-site EEG studies, personnel needs extend far beyond the research staff needed for data collection. In the CNTRACS project, we dramatically underestimated the amount of time required for these activities; it was not until data collection was underway that we realized the massive time commitment that would be required from the EEG supervisory team and research staff. In retrospect, we estimate that for the five sites, 260 participants, and 1560 datasets involved in the CNTRACS study, 50% effort from two junior faculty members and 100% effort from two advanced trainees (e.g., an advanced graduate student or postdoctoral fellow) with EEG experience over the 2-year period of data collection would have been sufficient for the EEG supervisory team. We found that data collection demands were covered by 10 members of the data collection team (2 at each site), and 9 members of the data preprocessing team were needed to complete preprocessing steps on all 1560 datasets over a 10-month period.

2.1.4 | Summary and Recommendations: Establish your teams

1. During the planning and budgeting phase, calculate how many hours will be needed to train staff, troubleshoot problems, process the data, conduct quality control meetings, summarize progress to the principal investigator(s), preprocess and analyze data, and any other elements of your data collection and monitoring protocol. Make sure that you have budgeted sufficient time for your data collection, data preprocessing, and EEG supervisory teams to meet these demands.
2. Consider including advanced trainees in the EEG supervisory team. Tasks that are ideal for advanced trainees include conducting the quality control meetings, training and monitoring the preprocessing team and responding to urgent messages from the data collection team that require immediate but low-level troubleshooting. Make sure that the trainees leading these monitoring activities are well supported and closely supervised by faculty.
3. Tasks that are ideal for highly experienced investigators include completion of site visits in multi-site studies, deep inspection of EEG quality during the first few weeks of data collection, supervision of advanced trainees, and communicating with the team of principal investigators.

2.2 | Before data collection: Experiment design

2.2.1 | Multi-task studies

Some special experimental design features should be implemented if your study includes more than one task. First, all tasks should be programmed using the same experimental control package (e.g., PsychToolbox (Brainard & Vision, 1997), PsychoPy (Peirce et al., 2019)) and the same programming structure (e.g., format for entering subject ID numbers, timing approach). This may seem inefficient if it requires reprogramming existing tasks, but it will save time in the long run: if and when errors in a script are discovered, this approach will make troubleshooting much easier. Additionally, this will lead to fewer errors in starting the scripts or entering relevant input data. Standardized scripting becomes doubly important when your study involves multiple testing sites. When problems arise, EEG supervisors can more easily identify them and guide research staff in implementing the solution remotely.

Second, a consistent "look and feel" should be implemented for the task demonstrations, practice, and prompts to begin recording. This approach benefits both the data collection team, who will have an easier time learning the flow of the tasks, and the participants, who will have an easier time digesting the task instructions. Additionally, we recommend using a prompt screen at the appropriate time to remind research staff to begin recording and the convention for naming the EEG file. The effort required to follow these suggestions is minor compared to the time required to deal with a failure to start recording, a filename that is not recognized by an analysis script, etc.

Third, ensure that your file naming conventions and data folder structures are optimal for the analysis approach that you will eventually perform, and minimize the opportunity for user error. This means that the analysis stream will need to be planned at the time of task creation (see below for data analysis and preprocessing pipeline considerations). For instance, you will want to ensure that both the behavioral and EEG file names include the following elements: participant ID number, task abbreviation, and version number. For behavioral files, additionally tagging the file name with a timestamp will ensure that behavioral data files are not accidentally overwritten. Altogether, the file name will look something like this: A001013_Task1_v2_03192022. Finally, if your participant ID number

includes leading letters and numbers (e.g., A00), consider writing your script such that the leading letters/numbers are auto-filled in the control prompt to avoid variation in experimenter procedures when starting the experiment.

Finally, a system should also be created for tracking versions of your experiment(s) and data analysis scripts (see *Processing Pipelines* for more details on analysis scripts). Multiple versions of stimulus presentation and data analysis scripts are inevitably generated in a large project as a result of bugs, pilot results, updates to operating systems, etc. If care is not taken, this can lead to significant problems when the wrong version of a script is used. We recommend using a centralized repository that is designed for tracking multiple versions, such as Github. Whichever system you use, it should include the following information:

1. A brief description of the task.
2. A detailed conceptual description of all script changes so that it is clear what updates were implemented in each version and why they were needed. If more than one version of a script is rolled out over the course of the study, document the date that it was implemented and tested at each data collection site.
3. All relevant task parameters, including the number of trials in each condition, stimulus timing information, trial structure, and a list of all event codes with their corresponding events.
4. Instructions for how to make specific modifications to the script(s). For example, experiment control scripts must usually be modified to assign a unique port address for event codes, and data analysis scripts may need to be modified to work under different operating systems. Detailed instructions for how to make such changes will minimize the risk of overlooking critical steps when updating and disseminating scripts.

These details are at the forefront of one's mind while developing the project, but they fade from memory with surprising speed. When the time comes to perform final analyses and write the multiple manuscripts that arise from a large study, it is useful to have this information stored in a location that can be easily accessed by all members of the study team, rather than buried in the code or notes from project planning meetings.

2.2.2 | Summary and recommendations: Task design

1. If your study includes multiple experimental tasks, program them such that they all use the same experimental control package, have the same structure, and begin with a similar sequence of prompts, demonstrations, and practice.
2. Carefully consider the naming convention and folder structure for saving behavioral and EEG data files, ensuring that it can efficiently accommodate your preprocessing and data analysis plan. Once your study is underway, it can be very time-consuming for the individuals analyzing the data to accommodate suboptimal data organization.
3. If needed and if the budget allows, hire a dedicated programmer who can convert existing tasks into the same experimental control package. This can be more cost-effective than it may seem and is worth considering if your existing team does not include a member with the necessary programming skills. For example, task programming or modification can often be accomplished by contracting advanced undergraduate or graduate students with good programming skills who may be contracted on an hourly basis.
4. Store each version of the task, detailed descriptions of task modifications, and experimental parameters in a centralized repository for easy access by all members of the investigatory team. For multi-site studies, it is especially important to keep track of the date that the script was officially updated and tested for errors at each site; in our experience, it is critical to obtain explicit confirmation from each site that updates have been implemented and tested to avoid a circumstance in which multiple versions of a given task are in use. Adding a "version" field to the behavioral data file or adding the version number to the file name is strongly recommended for minimizing the risk of such errors.

2.3 | Before data collection: Piloting

If you are using a new paradigm or a paradigm that has been meaningfully modified from a previous research study, begin by running a pilot study in a group of easy-to-test participants (e.g., 20 college students). In doing so, you can confirm that the paradigm works as planned before you invest effort, time, and money into data collection and processing. Even if piloting causes a short delay in beginning your study, the time cost is minor compared to the amount of time needed to correct errors after data collection has begun.

When piloting your tasks, you should ask several questions (see Tully and Boudewyn (2018) for a more complete overview):

1. Do the experiment instructions make sense to a naïve participant? Solicit feedback from your research assistants, who will be familiar with how the pilot participants react to the task.

2. Do you need a mechanism for assessing task comprehension, such as a short quiz that the participant takes before beginning the experiment, or an accuracy calculation from the practice? Should the participant be allowed to repeat the practice if they are not able to achieve satisfactory accuracy?
3. Are the event codes sensible and complete? Have you ensured that you will be able to test all of your planned comparisons? Do you have enough trials in each condition? We recommend ensuring that event codes are included for everything that occurs during an experiment. One benefit of large-scale studies is that they can often be re-analyzed in different ways—sometimes many years later—and it is not always possible to anticipate what information will be needed. As such, it is beneficial to err on the side of marking the onset of every event, even if it does not seem important for the planned analyses at the beginning of the study.
4. Do you see the expected number of event codes for each condition in your output file (see *Processing Pipelines*)? Verifying that number of codes per condition is exactly (not just approximately) correct often makes it possible to identify bugs or hardware problems.
5. If your task requires button-press responses, are the response event codes showing up properly?
6. Is your task too long? Can a participant reasonably complete the task in one session without becoming too restless, which will impact data quality? Are there enough breaks?
7. Is the behavioral data file saving properly? Can all planned behavioral analyses be performed with the data? Can the expected behavioral effects be observed in the pilot data? The behavioral data file should include all information needed to reconstruct the details of each trial of an experiment.

In addition, we strongly suggest that all anticipated analyses be conducted on the pilot data, including behavioral analyses (see *Processing Pipelines*, below). This is a challenge in practice because it can be painful to write all the analysis code before beginning data collection. However, you will need to write the analysis code eventually, and doing it at this early stage will allow you to find and resolve many potential problems while they can still be easily fixed. For example, you may discover that your event codes do not sufficiently differentiate trial types for some of your planned comparisons. Alternatively, you may discover that there is a problem with the timing of your trials such that there are stimulus overlap artifacts from the previous trial in your baseline. It is far better to discover these types of errors during the piloting phase rather than after data collection is underway. For multi-site studies, a minimum of one pilot subject should be run at each site following the finalization of the formal pilot phase, and the full analysis pipeline should be applied to each of the resulting data files. Multi-site studies may also want to consider running the same pilot subject at all sites, in order to be able to compare pilot data from the same participant across sites (although this can be costly, and there is a lack of an obvious standard for assessing similarity). Piloting and fully analyzing the pilot data serves the purpose of ensuring that no site-specific errors have been introduced (e.g., software version incompatibility, misidentified event code port, etc.), as well as providing an opportunity to inspect the data quality and make sure that it is satisfactory.

Finally, if you are studying a special population such as small children or a clinical sample, pilot your task in at least a few participants who are representative of your experimental sample to ensure that the task parameters are feasible. You may find, for example, that stimulus durations that were sufficient for healthy young adults are too short for individuals with processing speed impairments. Alternatively, you might find that task instructions that are understandable for college undergraduates need modification for individuals with less experience with computers (see also Kappenman and Luck (2016) for a discussion on considerations for data quality in clinical populations). Piloting your task to representative members of your target sample and soliciting feedback will reveal any such need for adjustments and help to ensure that your effects are interpretable once data collection is complete.

2.3.1 | Summary and recommendations: Piloting

1. Before beginning formal data collection, pilot your task with a sample of easy-to-access participants.
2. Use this pilot data to develop your EEG *and* behavioral data analysis pipelines, and ensure that you can perform all anticipated analyses with the event code structure and behavioral file format you have selected.
3. Pilot all tasks in at least one participant at every site in the case of multi-site EEG studies, and perform the entire analysis pipeline on each resulting data file to ensure that no site-specific errors have been introduced.
4. If your study includes special populations, ensure that task parameters are feasible for these participants by collecting a small amount of pilot data from members of your target sample.

2.4 | Before data collection: Test for event code delays

Almost all current video displays interpose a constant delay between the time the video card sends an image (which is when the event code occurs) and the time when the image appears on the screen. This delay varies across models, and it can be as long as 50 ms. Delays may also occur for stimuli in other modalities. In addition, if the stimulus presentation script is not written properly, a random delay may be added to this constant delay (see Chapter 16 in Luck (2014), for a detailed discussion). Whether your study is a single- or multi-site study, the event code delay should be measured at each site prior to the beginning of the study. For visual stimuli, this is accomplished by placing a photosensor in front of the display and recording the light emitted by your display for each stimulus (contact your EEG system's manufacturer for instructions). Measuring the delay is important for two reasons: first, you will need to shift the event codes in your analysis pipeline to account for the constant delay. Second, if you discover that you also have a variable delay, this means there is a bug in the stimulus presentation system that must be fixed. Even if the variable delay is small, it may indicate a bug with other significant consequences. In either case, you will want to be aware of such delays prior to undertaking a large data collection effort.

2.4.1 | Summary and recommendations: Test for event code delays

1. Prior to beginning data collection, measure the event code delay for each acquisition system and determine whether it is constant or variable.
2. Assuming a constant event code delay, it is usually sufficient to simply adjust the event code timing by the average delay across all trials in your analysis pipeline. For multi-site studies, this constant will likely differ by site.
3. If you discover that the timing of your event codes exhibits significant variability (i.e., more than ±1 sample period), this usually indicates a bug that must be eliminated before data collection begins.

2.5 | Before data collection: Processing pipelines and protocol documents

2.5.1 | Processing pipelines

It is essential to develop a few different processing pipelines during your pilot testing phase, each of which serves a different purpose. You will, of course, need your formal analysis pipeline to process the data for publication, and this should be developed during the piloting phase to ensure that you have everything you need to conduct the desired analyses (e.g., the appropriate event codes). However, it is also important to develop additional scripts to provide the research assistants and the supervisory team with feedback about data quality. In our experience, a few days of training prior to the beginning of data collection is not nearly enough for a research assistant to become truly proficient at collecting clean EEG data. Rather, they need frequent and meaningful feedback about the quality of the data they are collecting. We, therefore, recommend developing two scripts to provide this feedback.

The first is an initial quality assurance script that is executed immediately after each data collection session. This script performs simple, automated data cleaning and prints out basic quality metrics that provide immediate feedback for the research assistant who collected the data. The goal of this script is to provide simple but immediate feedback, while the recording session is fresh in mind. The second processing pipeline is a more sophisticated quality assurance script that research assistants can use to inspect their data in preparation for their regular meetings with the EEG supervisory team. Using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014), this script prompts the research assistants to visually inspect the data and observe the impact of the artifacts on an averaged ERP waveform. This helps the research assistants understand how their efforts to monitor and eliminate artifacts during the recording sessions impact the data. We find that this increases their motivation to be diligent during the electrode application procedure and pay close attention to the EEG during the recording sessions so that they can detect and solve problems that might otherwise render the EEG data unusable.

The elements of these pipelines and examples of how to structure them will be discussed in much more detail in Section 2; however, we preview this issue here because it is strongly recommended that you use your piloting process to develop these pipelines. If there are problems with the analyses you plan to perform, an unexpected number of trials, errors in the event codes, or unacceptably noisy data, you will be able to catch and correct it at this stage.

2.5.2 | Protocol documents

As with processing pipelines, there are at least two types of protocol documents that you may use, each of which serves

a different purpose. First, we recommend creating a long-form training manual that provides a detailed explanation of EEG setup procedures, suggestions for troubleshooting problems that arise, and examples of common artifacts and how to correct them. In our own long-form protocol document, we also include some text describing basic EEG theory and suggest that all members of the data collection team watch open-source instructional videos to familiarize themselves with common concepts (https://erpinfo.org/resources). The long-form protocol document should be reviewed with all new team members prior to beginning data collection. This can then be turned into a published protocol (typically without peer review) that can be cited in any publications resulting from the project, increasing the transparency and reproducibility of the research (e.g., see (Farrens et al., 2019; Simmons & Luck, 2020)).

Second, we recommend developing a short-form protocol checklist (also referred to as a *run sheet*) that is used to mark off each step of data collection for individual research participants. This document serves to ensure that all steps of the EEG setup protocol are performed and to minimize drift in setup practices over time as research assistants become more comfortable with the procedures. Importantly, this short-form checklist can also serve as a place to document any recording errors that arise. Examples of both the long-form training manual and the short-form protocol checklist are provided at https://osf.io/wdrj3/.

The actual use of these protocol documents is described in Section 2: *Create and Consistently Use a Runsheet and Session Notes*. However, we raise this issue here because these documents are best developed during the piloting phase and prior to training any new research personnel (see this section: *Personnel Training*). Solicit feedback from your research assistants regarding the use of the short-form checklist, and add any items or reminders that they indicate would be helpful.

Finally, investigators may consider recording training videos to accompany the protocol documents. This is particularly useful if the study is a multi-site or longitudinal study in which multiple data collection teams are involved or significant staff turnover is anticipated. However, it is also useful for smaller studies when similar protocols are used across many different studies. These videos can include demonstrations of participant greeting and consent, EEG cap placement, data collection, and clean-up protocols. For multi-site studies, we recommend that a single site record the components of the process from start to finish and disseminate the training videos to the other sites. For single-site studies, we recommend that these be used to build a repository of resources for training personnel in standard lab practice.

2.5.3 | Equipment considerations for multi-site studies

If your study involves more than one testing site, we strongly suggest that a senior member of the EEG supervisory team visits each site prior to data collection to set up equipment and software, even if sites already have existing EEG systems that will be used for the study. You will want to ensure consistency across sites in the equipment connections, software, response devices, sound systems, visual output from computer monitors, acoustic outputs, stimulus timing, and so on. Although this may seem like excessive effort, particularly for sites with established EEG data collection teams, it is essential to have the same person inspect every acquisition system. It is surprisingly easy to miscommunicate about equipment configurations, resources, and stimulus quality without this step.

If possible, all data collection sites should use identical stimulus presentation and EEG acquisition equipment and software (which can usually be justified in the grant budget for a large study). Although you might assume that minor differences in equipment across sites will be easy to accommodate in your processing pipeline, small discrepancies tend to compound as the complexity of the study grows, eventually incurring a significant time commitment by senior personnel. Moreover, reviewers might be concerned about the effects of these minor differences. Thus, it is worth the cost and effort to ensure that systems are as uniform as possible across sites.

If this uniformity is not possible, we recommend: (a) using a high sampling rate during data acquisition so that all systems can be downsampled to the same rate during preprocessing; (b) applying minimal filtering during data collection so that identical offline filters can be applied to the data from all systems during preprocessing; (c) using identical electrode montages; (d) recording calibration signals from each system during the piloting phase and ensuring that the signals are equivalent after passing through your processing pipeline.

For multi-site studies in which a new EEG system must be assembled at one of the data collection sites, we have found that it is more efficient for the local staff at a given site to assemble the system prior to the site visit by the EEG supervisor. The EEG supervisor can hold virtual meetings with the on-site staff to guide them through the assembly process and initial testing to make sure that the system is working properly. This ensures that the EEG supervisor can focus on inspecting the system and training the staff during the site visit and will not have to come back for a second visit because a critical piece of equipment was damaged during shipping or because a hard-to-find connector or cable was not purchased in advance.

2.5.4 | Summary and recommendations: Processing pipeline and protocol documents

1. All elements of the acquisition system should be as similar as possible across data collection sites unless otherwise indicated by the study design.
2. For multi-site studies, a member of the EEG supervisory team should visit each site and ensure acquisition systems are properly connected and calibrated. A checklist of all equipment needed should be used for each site visit (see example at https://osf.io/wdrj3/), and any exceptions should be documented so that they can be accounted for during preprocessing. For longer studies (i.e., those longer than a year), revisits by the EEG supervisory team should occur annually to ensure ongoing harmonization.
3. If the assembly of a new EEG system is required, local staff should assemble and test the system using virtual meetings with the EEG supervisor so that the supervisor can make the most of the limited time available during the site visit.

2.6 | Before data collection: Software

It is natural to think that you will only need to set up the hardware and software once at the beginning of the study. However, software or hardware changes are likely over the course of a large study because of situations like the following: (a) a computer fails unexpectedly and a new one must be equipped with all of the required software; (b) an operating system must be updated, which then causes the software to fail; (c) personnel changes necessitate new installations of software under different usernames. Without advance planning, these situations can cause significant delays in data collection, failed sessions, or corrupt data files. We suggest developing a formal plan for testing updated software and hardware, which will be facilitated by a comprehensive list of all software required to run the study (for more on mid-study troubleshooting, see Section 2: *Prepare for the Unexpected*). An example software checklist is included at https://osf.io/wdrj3/.

Video tutorials can be very useful for guiding research staff in software installation. Seeing a video of the installation process is typically much clearer than a text-based instruction such as "Open Matlab and set the path to include the folder where your analysis pipeline scripts are stored." These tutorials take time to create, but they can save a lot of time in the long run and minimize the potential for errors that result in corrupt or invalid data. As an example, if you implement the quality control analyses described above, and your data collection team involves 15 research assistants over the multiple years of your study, this is at least 15 times that you will need to ensure that EEGLAB and ERPLAB have been installed properly. It is far more efficient to develop a video tutorial rather than to schedule 15 separate meetings. An example video tutorial for downloading and installing EEGLAB and ERPLAB can be found at https://osf.io/wdrj3/. Note that, once these videos have been developed, they can often be used for many studies, making them even more efficient.

2.6.1 | Summary and recommendations: Software

1. Use the planning and piloting phase to develop a comprehensive list of all required software and their versions; use this list to ensure that all data collection systems are using identical software (unless otherwise indicated by the study design) and keep it as a reference for interruptions in the study that necessitate new installations.
2. Video tutorials are especially useful when several installations of each program will be needed.

2.7 | Before data collection: Personnel training

Of all of the recommendations outlined in this work, careful training of research staff in proper data collection procedures is perhaps the most important. These individuals are responsible for ensuring high-quality data collection, and when properly trained, they can rapidly detect errors that require immediate correction (e.g., event codes not appearing as they should, broken electrodes, etc.). As this role is so important, we discuss personnel training twice: In Section 2, we discuss recommendations for assisting research staff in maintaining good attention to detail after a study has begun; here, we discuss recommendations for initial training of new research staff. Even if your staff member arrives with EEG data collection experience from a previous lab, we still recommend asking them to complete the formal training to ensure that their data collection procedures are fully aligned with the study protocol. This is especially important in the case of multi-site studies, which may or may not have on-site EEG data collection expertise; in these cases, a visiting EEG supervisor may only have 2 or 3 days to ensure that all staff are fully trained and ready to begin. For this reason, *efficiency* in training is often of utmost importance.

A standardized, written personnel training plan should be developed during the preparation and piloting phase. This is essential for obtaining uniformity in data acquisition procedures, even if all data are collected at a single

site. In addition, a written plan will help maintain this uniformity when staff turnover occurs in the middle of data collection, which is likely in large-scale EEG studies. In the absence of such a plan, data collection practices are likely to drift at these junctures (think about how easily and quickly a message becomes distorted during a game of "telephone") and the consequences of failing to attend to the protocol can be significant (e.g., Kappenman and Luck (2010)). We have included an example from the CNTRACS study at https://osf.io/wdrj3/. Because some of the CNTRACS data collection sites did not have much local EEG expertise, this training plan is quite thorough and includes some of the elements described earlier in the context of the EEG supervisor's site visit (e.g., equipment and software checklists).

2.7.1 | Summary and recommendations: Personnel training

1. A formal training plan should be developed during the preparation and piloting phase. The most important and yet most difficult aspect of training new staff is ensuring that they follow the steps of the protocol as they are written to prevent drift in data quality over time or across data collection sites.
2. During training, emphasize to research staff that they should be messaging the EEG supervisory team in real-time when issues arise that need to be addressed immediately (see *Keeping Lines of Communication Open*). Building this habit into lab practices early is an excellent way to reinforce the message to research staff that they are expected to monitor continuously for errors and take appropriate steps immediately when a problem arises.
3. Ensure that all data collection staff members are aware of how to contact the EEG supervisory team for urgent questions, back up their EEG data, run the appropriate quality control analyses, and document any recording issues within the short-form protocol or run sheet.

2.8 | Before data collection: Keeping lines of communication open

During the planning stage, a detailed communication plan should be developed so that all members of the data collection team have an easy and reliable way to communicate with members of the EEG supervisory team when urgent issues arise during recording. Experienced members of the data collection team may be able to recognize artifacts in the data, but they may not have a good sense of which artifacts can be corrected offline and which artifacts must be eliminated during data collection. Sometimes, research staff are simply not sure if what they are seeing "looks right," for one reason or another. These issues are easily solved and data loss is prevented when research staff feel empowered to send a picture of the recording issue to the EEG supervisory team and know that they can expect a prompt response. This practice also reinforces the message that it is the data collection staff's responsibility to carefully monitor the data during the recording sessions and reach out for help when a potential problem is discovered. More than once during the CNTRACS project, a participant's data were saved by such timely interventions!

Oftentimes a simple format such as a group text is sufficient for the purposes of these urgent communications. As long as one member of the EEG supervisory team is available, research assistants can depend upon getting a prompt response. Alternately, it may be useful to use messaging applications such as *Slack* or *Microsoft Teams*. If all research staff within a study are part of the same messaging channel, everyone can view the recording issue and benefit from seeing the response from the EEG supervisors. As they gain more experience, they can even contribute their own suggestions for resolving the recording problem, thus decreasing the burden on EEG supervisors.

2.9 | Before data collection: Closing remarks

Whether your project is a single EEG study, a longitudinal study with several follow-up sessions over a period of years, or a multi-site consortium study, careful planning at the development stage can save substantial time and money over the course of the project and increase the likelihood that the large investment you make in collecting the data will pay off in robust, statistically significant findings. Although the impulse to move quickly and get the study launched is understandable, skipping some of these planning steps is likely to slow the completion of the study and could compromise data quality. Idiosyncrasies and hassles that appear minor and easy to accommodate at the beginning are easily forgotten or lost in the shuffle once the study has been underway for some time. Given the scope of these considerations, such planning ideally begins at the budgeting phase when writing your proposal; without a clear plan it is easy to underbudget on personnel needs or propose an unrealistic timeline.

The suggestions described thus far are intended to assist you with laying a strong foundation for launching your EEG study, irrespective of size and scope. However, even the most carefully planned study is likely to be thrown the occasional curveball after data collection has begun. In Section 2, we suggest several monitoring strategies for

ensuring high-quality data collection throughout your study, and for quickly catching and addressing any errors that arise.

3 | SECTION 2: WHAT TO DO ONCE DATA COLLECTION HAS BEGUN

Even with thorough and detailed planning before beginning a study, it is common for research practices to drift away from the original experimental protocol or for research personnel to develop idiosyncratic assessments of data quality over time. Consequently, it is important to establish clear and well-documented metrics of data quality and to monitor data quality continually over the course of the project. In addition, unexpected issues can arise over the course of a study; as described in Section 1, new rules, procedures, and code that are hastily concocted on the fly to fix unexpected problems can create a data collection procedure or analysis pipeline that is very difficult to later describe, replicate, or even remember. Our recommendations below are designed to help you avoid this situation.

3.1 | Data collection has started: Runsheets and session notes

We suggest developing a short-form protocol checklist, or *run sheet*, which includes a brief checklist of all steps to be followed during data acquisition as well as a place to take notes. As noted in Section 1, you should create this document during the piloting and planning phase, so that it is ready to use during the very first EEG recording session and at each and every data collection session after that. The run sheet should include a short-form checklist of the key elements of the experimental protocol (e.g., see: https://osf.io/wdrj3/), as well as a place for the in-the-moment notes that the person collecting the data takes during data acquisition. Completing the checklist for every data acquisition session will help ensure that all members of the research team follow the same list of steps and that no one ends up unintentionally creating their own version of set-up procedures. The run sheet notes can include everything from a short description of electrode connection issues that may come up during a session and how they are resolved, to a comment that a given participant seems to struggle with the task and needs a lot of breaks. Even if the notes amount to a brief statement that "everything went smoothly", the run sheet and notes section should still be completed. Completing the notes file reminds the data collection team that they should be monitoring throughout the recording and commenting about how things are going.

We recommend that the run sheet be saved and stored in the same folder as the raw data. In our sample data and scripts, we have saved the run sheet in the raw data folder in PDF format, such that it can be easily loaded during preprocessing and used to facilitate data inspection. The notes are an invaluable source of information during the data analysis stage of the project, and storing the notes in a file alongside the data ensures that they are easily accessible. As one example, the run sheet notes serve as a guide for marking bad channels during preprocessing. Additionally, two data files may exist for some participants (e.g., if the session was interrupted in the middle), and the notes are important for explaining why two files exist and how they should be handled. A given run sheet for a dataset might say something like "Electrode FC6 was giving us trouble throughout," or, "Participant appeared to misunderstand task instructions and so the task was re-started at 2:15 min." This information can be essential for appropriate data analyses, which may occur months or years after the data were collected.

We also suggest that the run sheet include reminders to upload and back up data files after each session. Our general rule is that each participant's data should always be available in at least two places, one of which is online (e.g., Box). Ideally, the "clean-up" protocol following each recording session would include checklist items for ensuring that the local copy of the data is saved appropriately and that the backup copy is uploaded.

Although run sheets are useful for all EEG studies, they are particularly important for large-scale studies in which the data preprocessing and analysis are not typically performed by the same individual who collected the data, or in which a large amount of time has passed between data collection and analysis. In these cases, the run sheet might be the only source of information available at the analysis stage about how things went during data collection. A sample run sheet can be found at https://osf.io/wdrj3/.

> ***When to Complete***: During data collection, in real-time.
> ***Who Completes This***: Person collecting the data.
> ***Suggested Form***: Digital document stored along with individual participant EEG data files.

3.2 | Data collection has started: Quality control assessments

During the piloting and planning phase, we suggest developing a series of quality control pipelines to monitor data collection and ensure that all data collection team members

continue to adhere to the same standards over time. In this section, we provide descriptions of each recommended quality control check. We also provide samples and resources including video tutorials at https://osf.io/wdrj3/.

A note about assessing EEG data quality: it is quite difficult to objectively assess, and as of this writing, there is no single, well-validated measure that is commonly used to determine the quality of EEG data (Clayson et al., 2021; Luck et al., 2021). With this in mind, our recommended approach is to use a series of checks that provide multiple opportunities to identify common problems at different stages from data collection to analysis. The suggestions provided below are not exhaustive, however, and investigators may wish to include additional data quality measures when adapting these guidelines for use in their own studies. The quality control assessments suggested below reflect those developed for the CNTRACS project that inspired this paper, but we ourselves would consider the inclusion of additional measures in future studies. For example, we have experimented with using automated preprocessing pipelines (e.g., PREP; Bigdely-Shamlo et al., 2015; HAPPE; (Gabard-Durnam et al., 2018; Monachino et al., 2021)) as part of the "First-Pass" Quality Control Check, so that EEG data quality measures such as the standardized measurement error (SME) can be computed as part of this initial quality control assessment. Thus, the reader should treat the suggestions below as a useful starting point for assessing quality control that could be improved upon as new measures are developed. For additional discussion of EEG data quality measures, please see Appendix 1.

3.2.1 | Pre-experiment check

We suggest that experimenters conduct a pre-experiment check of all tasks and equipment before each participant arrives. This can be as simple as opening up the task and ensuring that it runs as expected and that all software and hardware is working properly. It is even better to use some type of "dummy" subject (e.g., a calibration device or simply a resistor) to ensure that the recording hardware is working properly. While this cannot guarantee that nothing will go wrong during actual data collection (which is why we recommend the checks below), it can provide a valuable opportunity to troubleshoot or take steps to correct any problems before a participant arrives.

3.2.2 | "First-pass" quality control check

The first-pass quality control check is an abbreviated data processing script that is run immediately after each recording session to provide immediate feedback to the data collection team. We recommend that this include, at a minimum, (1) the number of occurrences of each distinct event code (aka trigger code) and (2) a summary of the participant's accuracy on the task. The output of the first-pass quality control check serves two functions. First, it provides an opportunity for major problems and unexpected errors to be detected that may have been missed by the individuals who are collecting the data. One of the most common causes of major data loss is a failure of transmission of event codes from the stimulus presentation system to the EEG recording system (as a result of an error in the task script or a hardware malfunction). Ideally, this would be discovered quickly at the beginning of the session and corrected. However, sometimes mistakes are made (not checking the event codes as they are coming out), or mistakes are difficult to spot (such as a small but important subset of event codes being lost or garbled). Completing a first-pass quality control check of the data immediately after recording it permits these mistakes to be caught as soon as they occur, and corrections can be made before additional participants are run. With a large-scale EEG project, this might be the only opportunity to catch a problem like this before many participants are run with the same problem, as a full analysis of all of the data may not occur for many months or even several years.

The second function of the first-pass quality control check is to remind the data collection team about the importance of continuously monitoring the data. Taking a few minutes right after an EEG session to run a short script like this reinforces the importance of collecting high-quality data, and keeps thoughts of things like event code tallying at the forefront of the data collection team's mind. We suggest that the data collection team be provided with specific guidelines for contacting the EEG supervisory team based on the results of this first-pass quality control check, such as the output displaying an incorrect number of event codes (see Section 1.8: *Keeping Lines of Communication Open*).

> ***When to complete***: Immediately after data collection (same day).
> ***Who completes this***: Person collecting the data.
> ***Suggested form***: Matlab script that is set up and confirmed to work as part of personnel training (See Section 1).

3.2.3 | In-depth quality control analysis

In large-scale and/or multi-site EEG studies, data collection and data analysis are often separated in terms of both time and personnel. For example, in the CNTRACS project that inspired this paper, ten full-time research assistants

across five different testing sites collected data that was later preprocessed by a different team of nine research assistants and then analyzed by the EEG supervisory team. In other studies, a rotating set undergraduate research assistants may collect the data over a period of years. Several years may pass before a "final" analysis of the data can be conducted with the full sample. You do not want to wait that long to discover problems with the data quality that could have been caught and corrected long before!

These problems include: (1) errors in the event codes or behavioral performance that made it past the initial data quality check but prevent critical comparisons from being made; (2) systematic issues with artifacts or data quality that would be obvious when performing a complete analysis and viewing the averaged ERPs. Having the data collection team complete the in-depth quality control analysis and attend regular meetings to review the data also helps to teach data collection team members about the different kinds of artifacts that can arise in the data, and the impact they can have on the analysis. This helps the data collection team members to hone their sense of what constitutes a critical issue in the data that needs immediate correction during acquisition, as opposed to minor artifacts that are unavoidable and/or not very problematic. This in turn leads to higher-quality recordings in the short run and greater statistical power in the long run.

An in-depth quality control analysis script should be developed during the piloting and planning phase, as recommended in Section 1. It should entail an abbreviated form of EEG preprocessing (excluding steps like artifact correction that require significant user input and experience), and it should produce a simple ERP plot to get a general sense of data quality. For example, if an experiment involves the presentation of any kind of visual stimulus, this script might involve basic segmenting, artifact rejection, filtering, and the creation of an average across all trials in all conditions to generate an ERP in which visual evoked potentials could be observed. Even if the goals of the experiment do not involve measuring the visually evoked potentials, this allows for the examination of the quality of the data based on the evaluation of an established, reliable ERP that should be observable in the data.

We suggest establishing weekly meetings with all members of the data collection team at the beginning of your study. Although you may plan to transition to a monthly schedule over the course of the project, weekly meetings should be conducted at the very beginning to establish good data monitoring practices. At these meetings, all data collection team members involved in data collection should come prepared to review the results of their in-depth quality control analysis, with a member of the EEG supervisory team to lead the meeting and answer any questions about how to handle artifacts observed in the data. The output of this analysis will include the basic ERP plots and results of the simplified artifact rejection routines included in the script. A sample script and dataset can be found at https://osf.io/wdrj3/.

In the early stages of a project, we suggest reserving time at each meeting to review all incoming data. As data collection team members become more experienced and the meetings move to a monthly schedule, priority can be given to data from sessions that there are questions about or was tricky in some way. We have found that this meeting lends itself well to a remote format (e.g., Zoom), in which each presenter shares their screen in turn to review the output produced by the in-depth quality control analysis. Discussion can focus on troubleshooting issues that can be spotted in the data, such as signal drift, "bad" electrodes, etc.

We have found it useful for the supervisor to provide feedback about the likelihood that the dataset being reviewed will yield usable data once processed. This binary feedback (likely usable vs. unlikely to be kept) can be extremely helpful to the data collection team members as they gain experience and learn about the kinds of artifacts that are relatively minor versus those that lead to datasets being excluded from the final analysis. They directly experience the time and effort required to run a session, so they are motivated to make sure that their sessions yield usable data. As discussed above, there is no agreed-upon "gold standard" objective measures of data quality, but the subjective judgment and feedback provided by the EEG supervisory team can serve as an important starting point for a group discussion on strategies for improving data quality in future sessions.

When to complete: On a rolling basis as data is collected, and presented as part of a supervised weekly meeting in which data collection team members summarize the results of this analysis for the group.
Who completes this: Data collection team member who collected the data.
Suggested form: We recommend that this analysis involve a "plug and play" script that requires little more than entering the participant ID number and hitting "run", once it is set up. By "set up", we mean that the script will have been developed prior to data collection, as recommended in Section 1, and that the data collection team members who will be collecting the data will have been assisted in installing it and trained on how to use it on their local computers.

3.3 | Data collection has started: Preprocessing pipeline

As discussed in Section 1, we strongly suggest that you develop the Processing Pipeline for your formal data

analysis during the piloting and planning phase—well before data collection begins. This pipeline can then be applied on a rolling basis once data collection begins. There are at least two ways to approach this: first, members of the EEG supervisory team can preprocess every file themselves. The advantage of this approach is that there is little or no need to do additional training to process the data; it is assumed that members of the EEG supervisory team have the requisite expertise to make decisions about things like channel interpolation and artifact correction. The disadvantage to this approach, of course, is that it is a labor-intensive task, especially for studies in which a large number of data sets is expected. This is true even if an automated pipeline is used (e.g., PREP; Bigdely-Shamlo et al., 2015; HAPPE; Gabard-Durnam et al., 2018; Monachino et al., 2021) because in practice the automated steps need to be verified and possibly tweaked by a trained researcher. Consequently, it may not be feasible for the EEG supervisors to complete all of the preprocessing for a large-scale study in addition to their other tasks. Thus, a second possibility is to invest in training a team of research assistants to preprocess the data. Although this approach can be time-consuming at the beginning, we have found that the invested time quickly pays off and is well worth the effort in the end. The suggested elements of this training approach are described below.

Before turning to our recommendations for training research team members on how to preprocess EEG datasets at scale, we want to include a note on the scope of this section. First, this section applies to standardizing the *pre*-processing pipeline. We define EEG preprocessing here as the set of steps involved in transforming raw data into a set of "clean" (i.e., free of bad channels and correctable artifacts) data files that are ready to undergo segmenting, artifact-rejection (of any artifacts still present in the data after preprocessing) and ultimately ERP averaging or time-frequency analysis. These latter steps are usually performed by a single member of the study team with sufficient expertise to finish the analysis and prepare the manuscript.

Our second note on the scope is that it takes time, training, and skill to analyze EEG data, and our goal here is not to provide a general background on EEG data analysis. There are very good existing resources available to guide new investigators through the process of getting started with data acquisition and analysis (Cohen, 2014; Farrens et al., 2019; Gable et al., 2022; Kappenman et al., 2021; Keil et al., 2014, 2022; Luck, 2014, 2022). There is also a variety of open-source code that is freely available with accompanying tutorials and sample datasets to learn how to do many of these steps (Cohen, 2014; Farrens et al., 2019; Gable et al., 2022; Kappenman et al., 2021; Luck, 2014, 2022). That said, there are some aspects of EEG preprocessing for which training is more difficult to obtain, and which involves making decisions about the data based on the judgment of the individual doing the analysis. Specifically, this applies to scanning raw data for "bad" channels and applying artifact correction methods like Independent Component Analysis (ICA) (Delorme et al., 2007). As such, training in these steps is often accomplished via a passed-down set of idiosyncratic lab-specific instructions that never make it into a tutorial or methods section. These unwritten rules can make it very difficult to establish reliability in the standards used by all members of the research teams.

Our suggested solution is to develop a standardized preprocessing pipeline coupled with thorough training for all data preprocessing team members who will be using it. Although there are some fully automated preprocessing pipelines that attempt to remove the subjective element from data screening and artifact correction (e.g., PREP; Bigdely-Shamlo et al., 2015), we recommend a semi-automated preprocessing pipeline (an example can be found at https://osf.io/wdrj3/). One reason is that, for the most part, the benchmark by which the quality of an automated preprocessing pipeline is judged is still the result obtained by a trained human analyzing the data. As has been noted elsewhere (Luck, 2014), humans are very good at the sort of thing these steps require—namely, pattern recognition. Another reason is that even if you choose to adopt a fully automated preprocessing pipeline, you will still need to have a look at the data in order to confirm that everything worked as intended (which requires the same kind of training approach advocated here).

We suggest developing a final analysis preprocessing script that yields clean, artifact-free data, but otherwise does not include many of the other processing steps that will need to be tailored to a specific analysis. This will allow you to flexibly use the product of the preprocessing scripts for many different analyses, without having to repeat the time-consuming data scanning and artifact correction steps. To accomplish this, we suggest that the product of the preprocessing script be un-segmented (continuous) data in which event codes have not yet been assigned to "bins" (condition categories). This will allow for later analysis-specific scripts to start by loading the same preprocessed data. Analysis steps such as segmenting the data into epochs and applying artifact rejection routines for any artifacts that remain in the data after preprocessing can then be applied.

Once your final analysis preprocessing script is in place, you will want to carefully train your data preprocessing team on how to use it. The goal is for all data preprocessing team members to arrive at the same conclusion concerning the identification of bad electrode channels and selection of ocular components/artifacts to remove in

a dataset. In other words, you want to minimize variability across individual members of the data preprocessing team in critical preprocessing decisions as well as drift in how a given individual preprocesses the data over time, both of which can lead to problems with interrater reliability and reproducibility. For example, even if all members of the data preprocessing team learn together how to arrive at the same preprocessing decisions, over time individuals can start to become more "lenient" or "strict" than others when evaluating data. Thus, after training, ongoing supervision and review of the data preprocessing team's results is critical when handling large-scale EEG datasets.

During the first phase of preprocessing training, we suggest providing all data preprocessing team members with background information on the types of artifacts that can be found in EEG data and the approach to correcting or rejecting them. Sample materials can be found at https://osf.io/wdrj3/. We also suggest that data preprocessing team members learn some conceptual EEG/ERP basics, such as by completing a free online course (e.g., https://courses.erpinfo.org/courses/Intro-to-ERPs). We then recommend that weekly training meetings be established, at which a member of the EEG supervisory team will first demonstrate how to use the script to preprocess data from one participant. We find that these meetings work well in a remote format (e.g., Zoom), such that screen sharing allows all data preprocessing team members to follow along with the preprocessing steps. Each data preprocessing team member then takes turns "leading" an analysis, such that all other members of the data preprocessing team can participate and help reach a consensus about data evaluation decisions. A sample training schedule and other materials including a Sample Preprocessing script and dataset can be found at https://osf.io/wdrj3/, along with detailed instructions and a tutorial video on how to use the script. Once reliability has been established (i.e., all data preprocessing team members arrive at the same selections for bad channel identification and artifact correction/rejection parameters), meetings can shift to a monthly schedule. We find it best if data preprocessing team members arrive at these monthly meetings prepared with their notes and questions about datasets they have processed since the last meeting, and with the data loaded and ready to review as a group.

When to complete: On a regular schedule (specific number of datasets to be processed in a given week, per preprocessor) once ~8 weeks of training has been completed; participation in monthly meetings thereafter to review results.
Who completes this: Data preprocessing team members.
Suggested form: We recommend using a script that generates plots and pop-up guideline reminders throughout the pipeline.

3.4 | Data collection has started: Prepare for the unexpected

3.4.1 | Turnover in research team members

Over the course of a large-scale or multi-site EEG study, as well as across smaller studies within the same laboratory, there will likely be a changeover in the research teams. Team members who are added after the study is underway will need to be trained, and it would be worth discussing at the outset of the study how this will be handled (see also a discussion of this in Section 1). When new data collection or data preprocessing team members join the project, will a member of the EEG supervisory team conduct a formal training, or will an experienced research assistant fill this role? If travel is involved (as in the case of a multi-site study), is there a budget for that? These are important questions to consider, ideally before data collection begins (See Section 1).

Simply repeating the initial training in experimental protocols that the original data collection team completed will probably not be enough to bring your new staff member up to speed. This is because innumerable small things will come up over the course of a project that have become unwritten rules and are stored only in collective lab memory. Here is a handful of examples of such things: knowledge of who in the department handles purchase orders for research supplies, and the procedure for making those purchases; the "trick" to getting the air conditioning to work properly; how to handle IRB modifications to an approved protocol; what all the acronyms used by a given lab mean; software license information, including passwords, user accounts (which might need to be transferred if someone leaves), and who can renew licenses when they expire; typical project-specific challenges to participant recruitment and how to address them; who to email for help when there is an unusual task error; the most efficient way to schedule the EEG room; the answers to common questions from participants during the consent process and EEG capping procedure; how best to organize the post-experiment clean-up tasks. We could go on. In our experience, losing a team member who knows all of those "unwritten" things and gaining one who must learn them through experience is the most difficult part of staff turnover. It is difficult not just for the principal investigators, but also for the new team member who is trying to get caught up.

The best solution to this problem is extensive documentation. It will never fully replace experience, but

detailed documentation can help speed up training new personnel, reduce stress for everyone involved, and remove many obstacles. As recommended in Section 1, we suggest establishing a long-form training manual at the beginning of the study. As the study progresses, the manual should evolve to contain updated information about the project and research teams, any changes to the experimental task and EEG recording protocol, procedures for ordering supplies, and any other aspect of day-to-day operations that are not documented elsewhere. Task the primary study coordinator at a given site with maintaining and adding to this manual as they go throughout the lifetime of the project. Did something change about the way the experimental software is accessed from the computer and loaded? Add a section describing this to the manual. Be specific! For example, do not just write that you "open the task". Provide the specific names of the files, the order they need to be opened, and what should happen when you click on them. Include screenshots so that the new team member can easily see exactly what is meant by the descriptions in the manual. Did the vendor for your EEG gel change? Add this to the manual. Is there a special procedure for downloading and installing the task that is only completed once and very easily forgotten after time goes by? Add it to the manual. You get the idea: document, document, document. When a research team member is preparing to move on from the lab, ask them to prioritize completing their additions to the project manual. Ideally, this will be a living document in which pieces of information like the above are consistently added as they develop.

3.4.2 | Task modifications

Our primary recommendation about making modifications to the task(s) for your large-scale study once data collection has already begun is to "just say no." Even small modifications to the script such as changing the task instructions or the number of trials introduce room for error because it is easy for multiple versions of a task to be in use. However, modifications are sometimes unavoidable, such as when critical errors in the script are discovered that would prevent the implementation of key data analyses, or it becomes clear that a participant group is struggling with the task at an unexpectedly high rate. In that case, use caution and keep in mind that your task modifications will mean that you will need good documentation about which participants were run on which version of the task, and when each version was implemented. The additional work and potential for error that this creates is the reason why extensive piloting and planning is emphasized in Section 1, and why our initial recommendation in this section is to avoid this situation altogether. If you must modify the experimental task(s), then unique version numbers should be created for every new rollout (beware of ever labeling things as "updated" or "final": this only leads to confusion later and file names like "Task_Final-Updated-Final-Final"). The version number should be included in the file-save prompt that appears when starting up an experimental task (as recommended in Section 1).

Remember that any changes that affect event codes or their numbers, timing, or trial structure will also like to involve the need for modifications to preprocessing and later analysis scripts for the lifetime of the dataset. Keep in mind also that every modification to the task that requires a modification to the processing pipelines will also require modification to all *copies* of all pipelines (and remember that some of these, such as the First-Pass Quality Control Check, are scripts on individual local computers) and monitoring to ensure that the changes were enacted on the proper datasets. As described in Section 1: *Experimental Design*, we recommend using a formal software versioning system (e.g., GitHub; Blischak et al., 2016). At a minimum, we suggest using a tracking sheet that documents each new version, its name, what changes were included from the last version, and the date this version was implemented. Ideally, a member of the EEG supervisory team would oversee the installation and documentation of task and script modifications.

4 | SECTION 3: CLOSING REMARKS

Large-scale EEG studies cost a lot of money and involve the allocation of many person-hours of labor, but they have the potential to make transformative contributions to science. When planning such a study, it is easy to focus only on the scientific questions, the experimental paradigms, and the predicted results. However, the collection of high-quality data is essential for the investment of money and time to pay off with replicable and high-impact contributions to science. If the EEG recordings are not clean or the preprocessing is not performed carefully and consistently, it will be difficult to obtain results that are statistically significant and reach conclusions that are scientifically sound.

It should be clear from this paper that a great deal of preparation, training, and monitoring are necessary to conduct EEG studies in a manner that is efficient and allows the investment of time and money to pay off. We hope that our experiences with the CNTRACS Consortium—including both our successes in advance planning and what we learned along the way—will allow others to conduct large-scale EEG studies that are efficient and ultimately produce major advances in science.

AUTHOR CONTRIBUTIONS

Megan A. Boudewyn: Conceptualization; writing – original draft; writing – review and editing. **Molly A. Erickson:** Conceptualization; writing – original draft; writing – review and editing. **Kurt Winsler:** Writing – review and editing. **John Daniel Ragland:** Writing – review and editing. **Andrew Yonelinas:** Writing – review and editing. **Michael Frank:** Writing – review and editing. **Steven Silverstein:** Writing – review and editing. **Jim Gold:** Writing – review and editing. **Angus W. MacDonald:** Writing – review and editing. **Cameron S. Carter:** Writing – review and editing. **Deanna M. Barch:** Writing – review and editing. **Steven J. Luck:** Conceptualization; writing – original draft; writing – review and editing.

ORCID

Megan A. Boudewyn https://orcid.org/0000-0001-7948-6303
Molly A. Erickson https://orcid.org/0000-0002-5311-4402
Steven J. Luck https://orcid.org/0000-0002-3725-1474

REFERENCES

Barch, D. M., Carter, C. S., Arnsten, A., Buchanan, R. W., Cohen, J. D., Geyer, M., Green, M. F., Krystal, J. H., Nuechterlein, K., & Robbins, T. (2009). Selecting paradigms from cognitive neuroscience for translation into use in clinical trials: Proceedings of the third CNTRICS meeting. *Schizophrenia Bulletin*, *35*(1), 109–114. https://doi.org/10.1093/schbul/sbn163

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*, 16. https://doi.org/10.3389/fninf.2015.00016

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A quick introduction to version control with Git and GitHub. *PLoS Computational Biology*, *12*(1), e1004668. https://doi.org/10.1371/journal.pcbi.1004668

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897x00357

Clayson, P. E., Brush, C. J., & Hajcak, G. (2021). Data quality and reliability metrics for event-related potentials (ERPs): The utility of subject-level reliability. *International Journal of Psychophysiology*, *165*, 121–136. https://doi.org/10.1016/j.ijpsycho.2021.04.004

Cohen, J. D., & Insel, T. R. (2008). Cognitive neuroscience and schizophrenia: Translational research in need of a translator. *Biological Psychiatry*, *64*(1), 2–3. https://doi.org/10.1016/j.biopsych.2008.04.031

Cohen, M. X. (2014). *Analyzing neural time series data: Theory and practice*. MIT Press. https://doi.org/10.7551/mitpress/9609.001.0001

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, *34*(4), 1443–1449. https://doi.org/10.1016/j.neuroimage.2006.11.004

Donoghue, T., Haller, M., Peterson, E. J., Varma, P., Sebastian, P., Gao, R., Noto, T., Lara, A. H., Wallis, J. D., & Knight, R. T. (2020). Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, *23*(12), 1655–1665. https://doi.org/10.1038/s41593-020-00744-x

Donoghue, T., Schaworonkow, N., & Voytek, B. (2022). Methodological considerations for studying neural oscillations. *European Journal of Neuroscience*, *55*(11–12), 3502–3527. https://doi.org/10.1111/ejn.15361

Farrens, J. L., Simmons, A. M., Luck, S. J., & Kappenman, E. S. (2019). *Electroencephalogram (EEG) recording protocol for cognitive and affective human neuroscience research*. https://doi.org/10.21203/rs.2.18328/v4

Gabard-Durnam, L. J., Mendez Leal, A. S., Wilkinson, C. L., & Levin, A. R. (2018). The Harvard automated processing pipeline for electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data. *Frontiers in Neuroscience*, *12*, 97. https://doi.org/10.3389/fnins.2018.00097

Gable, P., Miller, M., & Bernat, E. (2022). *The Oxford handbook of EEG frequency*. Oxford University Press. ISBN: 9780192653376, 0192653377.

Haumann, N. T., Parkkonen, L., Kliuchko, M., Vuust, P., & Brattico, E. (2016). Comparing the performance of popular MEG/EEG artifact correction methods in an evoked-response study. *Computational Intelligence and Neuroscience*, *2016*, e7489108. https://doi.org/10.1155/2016/7489108

He, B. J. (2014). Scale-free brain activity: Past, present, and future. *Trends in Cognitive Sciences*, *18*(9), 480–487. https://doi.org/10.1016/j.tics.2014.04.003

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*, 163–178. https://doi.org/10.1111/1469-8986.3720163

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *NeuroImage*, *225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, *47*, 888–904. https://doi.org/10.1111/j.1469-8986.2010.01009.x

Kappenman, E. S., & Luck, S. J. (2016). Best practices for event-related potential research in clinical populations. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(2), 110–115. https://doi.org/10.1016/j.bpsc.2015.11.007

Kappenman, E. S., MacNamara, A., & Proudfit, G. H. (2015). Electrocortical evidence for rapid allocation of attention to threat in the dot-probe task. *Social Cognitive and Affective Neuroscience*, *10*(4), 577–583. https://doi.org/10.1093/scan/nsu098

Keil, A., Bernat, E. M., Cohen, M. X., Ding, M., Fabiani, M., Gratton, G., Kappenman, E. S., Maris, E., Mathewson, K. E., & Ward, R. T. (2022). Recommendations and publication guidelines for

studies using frequency domain and time-frequency domain analyses of neural time series. *Psychophysiology*, *59*(5), e14052. https://doi.org/10.1111/psyp.14052

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., Luu, P., Miller, G. A., & Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*(1), 1–21. https://doi.org/10.1111/psyp.12147

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213. https://doi.org/10.3389/fnhum.2014.00213

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT Press. ISBN: 0-262-52585-2.

Luck, S. J. (2022). *Applied event-related potential data analysis*. LibreTexts. https://doi.org/10.18115/D5QG92

Luck, S. J., & Gold, J. M. (2008). The translation of cognitive paradigms for patient research. *Schizophrenia Bulletin*, *34*(4), 629–644. https://doi.org/10.1093/schbul/sbn036

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology*, *58*, e13793. https://doi.org/10.1111/psyp.13793

Monachino, A. D., Lopez, K. L., Pierce, L. J., & Gabard-Durnam, L. J. (2021). The HAPPE plus event-related (HAPPE+ ER) software: A standardized processing pipeline for event-related potential analyses. *BioRxiv*, 2021–07 https://doi.org/10.3389/fnins.2018.00097

Niso, G., Krol, L. R., Combrisson, E., Dubarry, A. S., Elliott, M. A., François, C., Héjja-Brichard, Y., Herbst, S. K., Jerbi, K., & Kovic, V. (2022). Good scientific practice in EEG and MEG research: Progress and perspectives. *NeuroImage*, *257*, 119056. https://doi.org/10.1016/j.neuroimage.2022.119056

Olvet, D. M., & Hajcak, G. (2009). Reliability of error-related brain activity. *Brain Research*, *1284*, 89–99. https://doi.org/10.1016/j.brainres.2009.05.079

Pavlov, Y. G., Adamian, N., Appelhoff, S., Arvaneh, M., Benwell, C. S. Y., Beste, C., Bland, A. R., Bradford, D. E., Bublatzky, F., Busch, N. A., Clayson, P. E., Cruse, D., Czeszumski, A., Dreber, A., Dumas, G., Ehinger, B., Ganis, G., He, X., Hinojosa, J. A., … Mushtaq, F. (2021). #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex*, *144*, 213–229. https://doi.org/10.1016/j.cortex.2021.03.013

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*, 127–152. https://doi.org/10.1111/1469-8986.3720127

Sadiya, S., Alhanai, T., & Ghassemi, M. M. (2021). Artifact detection and correction in EEG data: A review. 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER), 495–498 https://doi.org/10.1109/NER49283.2021.9441341

Simmons, A. M., & Luck, S. J. (2020). Protocol for reducing COVID-19 transmission risk in EEG research. *Research Square*, 1–14. https://doi.org/10.21203/rs.3.pex-974/v2

Tully, L. M., & Boudewyn, M. A. (2018). *Creating a novel experimental paradigm: A practical guide*. SAGE Publications Limited. https://doi.org/10.4135/9781526437495

Zhang, G., & Luck, S. J. (2022). Variations in ERP data quality across paradigms, participants, and scoring procedures. BioRxiv, 2022–08 https://doi.org/10.1101/2022.08.22.504647

How to cite this article: Boudewyn, M. A., Erickson, M. A., Winsler, K., Ragland, J. D., Yonelinas, A., Frank, M., Silverstein, S. M., Gold, J., MacDonald, A. W. III, Carter, C. S., Barch, D. M., & Luck, S. J. (2023). Managing EEG studies: How to prepare and what to do once data collection has begun. *Psychophysiology*, *00*, e14365. https://doi.org/10.1111/psyp.14365

APPENDIX 1

This appendix describes several metrics of data quality for EEG and ERPs. Although the signals of interest in EEG/ERP research are often tiny relative to the noise, which can dramatically reduce statistical power, the field has not converged on a standardized and widely used set of objective data quality metrics. Here, we discuss several metrics that could be valuable.

However, we would first like to make a fundamental but non-obvious point: Data quality must be defined with respect to the specific *score* that will be derived from the data (i.e., the dependent variable that will be entered into the statistical analyses), and there is no such thing as a universal, score-independent metric of EEG signal quality. For example, 50 or 60 Hz line noise can be a major problem when quantifying gamma-band activity in a time-frequency analysis or peak amplitude in the auditory brainstem response, but this noise has minimal impact when quantifying theta-band activity or measuring the mean amplitude of the P3 wave as the mean voltage between 400 and 600 ms. By contrast, low-frequency noise from skin potentials will greatly distort theta-band and P3 amplitude scores but will have little impact on gamma-band power or auditory brainstem responses (see Luck, 2022 for a detailed discussion). Thus, the data quality metrics for a given study should be tailored to the specific scores that are the focus of the study.

Raw EEG data quality

Some sources of noise have characteristic frequency-domain properties. For example, when a high-impedance

EEG recording system is used, skin potentials produce low-frequency noise (mainly less than 3 Hz) that can have a very large impact on statistical power for late components such as the P3b, N400, and late positive potential (Luck, 2022). Similarly, induced electrical noise is typically at a characteristic frequency (50 or 60 Hz), and muscle noise typically produces a notable increase in high-frequency activity (>20 Hz). Thus, depending on the details of a given study, the amplitude or power within a given frequency band could be a valuable metric of data quality for the continuous EEG.

Many software packages can provide this information. For example, ERPLAB Toolbox (Lopez-Calderon & Luck, 2014) includes a routine that quantifies the amplitude or power within a set of default or user-defined frequency bands (version 9.10 and higher). A threshold could be applied to this information in the data quality assessment scripts, and any dataset that exceeds the threshold would then be carefully examined to determine the cause of the noise.

Frequency-domain representations can also be used to quantify aperiodic noise. That is, when transformed into the frequency domain, aperiodic noise in the EEG typically creates a spectrum of power that falls off with frequency according to a 1/f function (He, 2014). The FOOOF toolbox can be used to quantify the amount of this aperiodic noise (Donoghue et al., 2020). This may be especially important in studies that focus on frequency-domain measures, because the aperiodic noise mixes with the true oscillations, degrading the ability to quantify the oscillations (Donoghue et al., 2022).

It can also be valuable to assess the presence of extreme values or flatline periods in the EEG, which may arise from head movements, scratching, amplifier saturation, etc. Algorithms for quantifying these can be found in EEGLAB Toolbox (Delorme & Makeig, 2004), ERPLAB Toolbox (Lopez-Calderon & Luck, 2014), and the PREP pipeline (Bigdely-Shamlo et al., 2015).

Number of channels and trials surviving exclusion

Another useful data quality metric is the amount of data from a given participant that will be excluded from the final analysis, which can be subdivided into "bad" channels and "bad" trials. Problems with channels are typically a result of a hardware problem (e.g., a broken electrode), a poor electrical connection with the skin, or bridging of nearby electrodes. For most studies, interpolation from the other electrodes is a satisfactory solution for bad channels as long as the number of bad channels is not too high. Thus, quality control scripts can report the number of bad channels and whether this number exceeds a threshold (where the threshold will depend on the nature of the study).

The most common problems with trials are behavioral errors (if correct responses will be required during the final analyses) and artifacts. Many studies use artifact correction algorithms (e.g., Haumann et al., 2016; Jung et al., 2000; Sadiya et al., 2021) to deal with the most common kinds of artifacts, such as eyeblinks. However, some types of artifacts are not easily corrected using these algorithms (e.g., movement artifacts and blinks that prevent the perception of a stimulus), so some trials will likely need to be rejected because of artifacts (Luck, 2014). If the number of trials excluded because of behavioral errors or artifacts is sufficiently high, the signal-to-noise ratio may be substantially reduced. Thus, quality control scripts can report the number of rejected trials and whether this number exceeds a threshold (where the threshold will depend on the nature of the study).

Final decisions about which channels and trials are "bad" are often made by experts during the final data analyses, after the study is complete. For example, in a study that follows artifact correction with artifact rejection, the number of rejected trials cannot be determined until after artifact correction, but artifact correction may be time-consuming and require considerable expertise. Consequently, definitive counts of the number of bad channels and excluded trials may be determined too late for the quality control procedures, which are designed to catch problems with data collection while the study is ongoing so that these problems can be fixed. A potential solution to this problem is to use automated routines that can provide preliminary estimates of the number of problematic channels and trials that are good enough for use by the quality control procedures that are run while a study is ongoing.

For example, the PREP pipeline (Bigdely-Shamlo et al., 2015) can be used to determine the number of bad channels and to perform artifact correction automatically, without requiring an expert. Automated artifact detection routines can then be applied to estimate the number of trials that will need to be excluded. An expert will ultimately be needed to verify that these automated processes have worked appropriately for each dataset, but completely automated processing should typically be sufficient for the purposes of the quality control procedures.

Baseline noise

A common informal approach to assessing data quality in averaged ERPs is to visually inspect the noise level during the baseline period (Luck, 2014), with a relatively flat baseline indicating low noise. This noise level (the lack of flatness) can be quantified by taking the standard deviation or root mean square amplitude of the voltages during the baseline period. However, two caveats to this

approach should be considered. First, this approach assumes that the noise level during the baseline period is a reasonable estimate of the noise level during the period of the ERP component of interest. This may not always be true. Second, this approach assumes that the kinds of noise that create millisecond-to-millisecond variation in the pre-stimulus voltage are the same kinds of noise that will degrade the ability to measure the amplitude or latency score of interest. However, measures of baseline noise will be relatively insensitive to low frequencies, and low-frequency noise is often the most significant threat to statistical power when late components such as P3b and N400 are being measured (Kappenman & Luck, 2010). Thus, although quantifying the baseline noise can help identify some types of noise during the quality control process, it should be supplemented by other metrics of data quality.

Metrics of reliability and precision

In most ERP studies, the single-trial EEG epochs are averaged to improve the signal-to-noise ratio, and the amplitude or latency of a given ERP component is then scored from the averaged ERP waveform. The resulting scores are the dependent variables that are entered into the statistical analyses and used to test the scientific questions of the study. Ultimately, the quality of the EEG data matters to the extent that it impacts these amplitude and latency scores. There are two main approaches to determining data quality in terms of these scores, one focusing on the *reliability* of the scores and one focusing on the *precision* of the scores (Niso et al., 2022).

Traditional psychometric approaches to reliability are based on correlations. For example, split-half reliability is computed by dividing the trials for each participant into two halves, obtaining the score of interest (e.g., P3b peak latency) from each half of the trials, and using the Pearson *r* correlation coefficient to determine the extent to which the two scores for each participant are correlated (see e.g., Kappenman et al., 2015; Olvet & Hajcak, 2009). The major benefit of this approach is that the resulting reliability value is directly related to the ability to detect correlations between the score (when obtained from all the trials) and other variables (e.g., the correlation between P3b peak latency and a measure of intelligence). The major downside of this approach in the context of quality control is that it yields a single value for the entire group of participants rather than providing a metric of data quality for each individual participant (but see Clayson et al. (2021) for a variant that can provide single-participant values in some situations). Thus, traditional reliability measures would not be practical for monitoring data quality while data collection is ongoing.

A more suitable approach for quality control would be to quantify the *precision* of a score, defined as the extent to which noise in the data would be expected to cause the observed score to differ from the true score (i.e., the score that would be obtained in the absence of noise). Specifically, Luck et al. (2021) proposed a new metric called the *standardized measurement error (SME)*, which is a variant of the more general concept of the *standard error of measurement*. The SME is computed separately for each participant, providing a subject-level metric of data quality. The SME can also be computed for complex scores, such as the onset latency of a difference wave. It could also be adapted for use with frequency-domain scores.

The SME is directly related to the expected effect size, so the SME indicates how the data quality for a given participant impacts the statistical power for the specific amplitude or latency scores being examined in a given study. The SME takes into account both the trial-to-trial variability and the number of trials that were available for averaging (e.g., after rejecting trials because of artifacts or incorrect behavioral responses). To quantify the trial-to-trial variability independently of the number of trials, it is possible to compute the standard deviation (SD) rather than the SME. Benchmark SD and SME values are available for seven widely-used ERP components (Zhang & Luck, 2022). The main downside of this approach at present is that it is not widely implemented in ERP analysis packages. However, it is included in versions 8 and higher of ERPLAB Toolbox (Lopez-Calderon & Luck, 2014), so it is freely available to anyone who is running Matlab.