International Journal of Psychophysiology 111 (2017) 57–67

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting

Peter E. Clayson [a,*], Gregory A. Miller [a,b]

[a] Department of Psychology, University of California, Los Angeles, Los Angeles, CA, United States
[b] Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA, United States

ARTICLE INFO

Article history:
Received 18 April 2016
Received in revised form 5 August 2016
Accepted 9 September 2016
Available online 10 September 2016

Keywords:
Event-related potentials
Psychometrics
Guidelines
Dependability
ERP reliability analysis toolbox

ABSTRACT

Failing to consider psychometric issues related to reliability and validity, differential deficits, and statistical power potentially undermines the conclusions of a study. In research using event-related brain potentials (ERPs), numerous contextual factors (population sampled, task, data recording, analysis pipeline, etc.) can impact the reliability of ERP scores. The present review considers the contextual factors that influence ERP score reliability and the downstream effects that reliability has on statistical analyses. Given the context-dependent nature of ERPs, it is recommended that ERP score reliability be formally assessed on a study-by-study basis. Recommended guidelines for ERP studies include 1) reporting the threshold of acceptable reliability and reliability estimates for observed scores, 2) specifying the approach used to estimate reliability, and 3) justifying how trial-count minima were chosen. A reliability threshold for internal consistency of at least 0.70 is recommended, and a threshold of 0.80 is preferred. The review also advocates the use of generalizability theory for estimating score dependability (the generalizability theory analog to reliability) as an improvement on classical test theory reliability estimates, suggesting that the latter is less well suited to ERP research. To facilitate the calculation and reporting of dependability estimates, an open-source Matlab program, the ERP Reliability Analysis Toolbox, is presented.

1. Introduction: measurement in psychophysiology

It is now widely understood that some of the most exciting work in psychopathology involves discovering and understanding relevant brain mechanisms, without falling prey to naïve reductionism (Lilienfeld, 2007; Miller, 1996, 2010). Despite the enthusiastic press such work receives, it is far from clear how to proceed. Many avenues beckon, and some of the most exciting research tools, being relatively new, are often the most primitive, demanding, and fickle. Some in the field have expressed alarm at the low likelihood that many currently appealing findings will stand the test of time. One of the challenges is the diversity of skills needed to understand, apply, and improve recent tools. A well prepared scholar has gathered a diverse set of skills and interests that allow a rich but cautious pursuit of biological mechanisms in psychopathology. Psychophysiology, especially hemodynamic and electromagnetic neuroimaging, has been greatly hampered by a widespread failure to consider a host of issues such as psychometric reliability, differential deficit, and statistical power. This oversight is a central contributor to recently rising concerns about replicability in this and many other areas of science.

Mismeasurement of psychological constructs can lead to misunderstood phenomena and mistaken conclusions that persist well after the publication of a study and that slow the progress and indeed the trustworthiness of psychological and biological science. Knowing how constructs are operationalized by a measurement method is paramount in understanding the applicability of that measurement to theory. Research using event-related brain potentials (ERPs) typically attempts to examine changes in neural activity correlated with various sensory, motor, cognitive, or emotional events to make inferences about psychological phenomena. Rigorous evaluation of the psychometric properties of ERP component scores is not commonplace in the literature. Without fidelity to continuous psychometric evaluation of ERP measurement, problematic inferences are inevitable, and the pace of psychological science is slowed.

The fundamental purpose of measurement in psychological science is to make inferences about psychological constructs based on observed scores that generalize to other samples, contexts, and outcomes. The ability to make such generalized inferences rests in part on the reliability and validity of the scores obtained from an instrument. The reliability of an ERP measurement captures the level of consistency or stability of that measurement, whereas the construct validity of an ERP measurement refers to the extent to which the measurement reflects the precise neural or psychological phenomenon it is intended to measure. Validity is concerned with the meaning and interpretation of a score obtained by

DOI of original article: http://dx.doi.org/10.1016/j.ijpsycho.2016.10.012.
* Corresponding author.
E-mail address: peter.clayson@gmail.com (P.E. Clayson).

http://dx.doi.org/10.1016/j.ijpsycho.2016.09.005
0167-8760/

a measurement and is not a property of the measure itself (Cronbach and Thorndike, 1971; Messick, 1989, 1995). Similarly, reliability is a property of scores, the data in hand, and not a property of a measure (Thompson, 2003; Vacha-Haase, 1998). Although reliability does not require validity, in classical test theory validity is limited by reliability. In order for an ERP measurement to be valid, it must be reliable. To draw solid conclusions about the psychology-biology relationships assessed by ERPs, the measurement approach used to quantify ERP components must first be demonstrated as reliable and valid in relevant contexts. This goal is particularly important as the Research Domain Criteria (RDoC) project of the US National Institute of Mental Health fosters expanded reliance on and development of continuous/dimensional measures in pursuit of hybrid psychological-biological constructs (Kozak and Cuthbert, 2016; Miller et al., 2016; Yee et al., 2015). A number of factors warrant consideration when designing, analyzing, and evaluating ERP studies. (The present paper focuses on ERP use, though some points apply throughout psychophysiology and beyond.)

2. Psychometric properties are context-dependent

Reliability and validity as properties of a measure are not universal but are dependent on a specific population and context, and they should be continually assessed and refined (Smith and McCarthy, 1995). Thus, a measure cannot be said to be reliable or valid in some general sense. It is commonplace to claim score reliability via citing previous psychometric studies (Vacha-Haase et al., 2000; Whittington, 1998), based on the common misunderstanding that reliability is an fundamental property of a measure (Vacha-Haase, 1998; Vacha-Haase et al., 1999), but score reliability and validity are context-dependent and cannot be assumed based on prior reports. For example, the reliability and validity values for a score on a questionnaire assessing depressive symptoms in an undergraduate sample cannot be assumed to generalize to psychiatric, developmental, and neurological populations or to outpatient clinics, inpatient units, community settings, or settings in other cultures. The clinical implications of a score – even the same score – on a questionnaire are likely to differ based on whether a student at a university or a patient in a hospital is completing the questionnaire. The reliability and validity of the measurement score often warrant evaluation in each sample and context. Doing so fosters measurement and effective operationalization. Since score reliability on a questionnaire can vary across administrations, even in similar contexts and populations, it is recommended that score reliability for questionnaires be reported in every study (Thompson and Snyder, 1998). Similar to a questionnaire, ERP score reliability and validity established in a specific context cannot be assumed to apply to other contexts and should be reported routinely.

The context-specific dependence of the reliability and validity of ERP scores can be readily observed in psychometric work on the error-related negativity (ERN), a scalp-recorded ERP that follows error commissions in speeded choice-response tasks (Falkenstein et al., 1991; Gehring et al., 1993; Larson et al., 2014). Studies examining the effect of population and context on ERN score reliability observe differential internal consistency depending on the clinical diagnosis being examined, such as psychosis (Foti et al., 2013) and anxiety or major depressive disorders (Baldwin et al., 2015). Furthermore, estimates of the number of trials needed to obtain acceptable levels of internal consistency (e.g., Cronbach's $\alpha > 0.70$) for ERN amplitude vary widely across studies, ranging from as few as two trials in healthy undergraduate subjects (Pontifex et al., 2010) to over 30 trials in a sample of participants with major depressive disorder (Baldwin et al., 2015). Given so much variability in observed reliability, if studies assume reliability based on previous psychometric work, interpretations based on ERN scores for which reliability was not demonstrated afresh would be in question. Since validity is often limited by reliability, and ERN score reliability is contingent on population, poor reliability also compromises the comparability of ERN interpretations examining different populations. Differences in neural processes manifesting as ERN may contribute to the context-specific reliability and could result in overgeneralization of population-specific relationships between ERN and other phenomena (see Meyer et al., 2013). Although the example just provided demonstrates the context-dependent nature of internal consistency, the same considerations apply to other types of reliability.

The task used to elicit ERPs is another source of variability that further contributes to the context-specific dependence of the reliability and validity of ERP scores. With regard to ERN, different estimates of internal consistency have been observed based on the task used to elicit it (Foti et al., 2013; Meyer et al., 2014; Meyer et al., 2013). ERN scores measured during the Flanker, Stroop, and Go/NoGo tasks show both shared and unique variance, suggesting a task-specific effect on observed ERN scores and providing evidence for convergent and divergent validity of ERN scores across tasks (Riesel et al., 2013). Divergent internal consistency and test-retest estimates of different ERP component scores are also observed within the same task and across different tasks (e.g., Cassidy et al., 2012), highlighting that ERP component scoring as a method does not produce consistently reliable results even when standardized procedures are followed (e.g., Keil et al., 2014).

Over and above the possible effects of the population sampled or the task used to elicit an ERP, the reliability and validity of ERP scores can be affected by diverse aspects of signal recording and processing (Edgar and Miller, 2016; Glaser and Ruchkin, 1976). The data-analysis pipeline starting with data collection and leading to an ERP score is extensive, variable, and dependent on the judgment of the researcher. Although consensus guidelines for data reduction for ERP analysis have long been available (Donchin et al., 1977; Keil et al., 2014; Picton et al., 2000), in practice many choices about signal recording and processing are left to the researcher, leading to numerous unique pipelines that surely affect the reliability and validity of ERP scores. A study of fMRI methods, which similarly entails numerous choices about recording parameters and processing steps, observed 223 unique data analysis pipelines among 241 studies, with many studies omitting details about data acquisition and analysis (Carp, 2012). Even though the ERP literature is much more mature, it likely has similar variability in data-analysis pipelines.

2.1. Noise is a critical challenge

The effect of different approaches to handling ERP noise, such as irrelevant physiological activity or environmental interference, demonstrates the impact that numerous choices can have on data recording and processing of ERP scores. Given the effect of noise on measurement error, which can be defined conceptually as fluctuation in scores that is irrelevant to the construct being measured, an ERP measurement is reliable only insofar as noise has been minimized in the waveform (Glaser and Ruchkin, 1976; Perry, 1966). Low measurement error is essential for good reliability (Nunnally and Bernstein, 1994), and ERP noise increases measurement error and decreases statistical power (Luck, 2014). The signal-to-noise ratio (SNR) for an averaged ERP can be understood using the formula: $(1/\sqrt{N})*R$, where N represents the number of trials included in the average, and R represents noise (Luck, 2014). In principle, the SNR increases as a function of the inverse of the square root of the number of trials included in the average. As in classical test theory, this principle assumes consistent signal (true score) and random error across trials, assumptions which are often doubtful.

Many aspects of the context in which data are recorded affect noise, such as physiological or environmental interference (Luck, 2014) or type of EEG system. For example, active electrodes may record data with a higher SNR at higher impedances than passive electrodes, although passive electrodes record the cleanest data at very low impedances (Laszlo et al., 2014; Mathewson et al., in press). High-electrode-impedance recordings may also be more susceptible to noise contamination from skin potentials in warm, humid recording environments (Kappenman and Luck, 2010). Although steps can be taken after