Ambulatory Assessment in Psychopathology Research:
A Review of Recommended Reporting Guidelines and Current Practices

Timothy J. Trull

University of Missouri

Ulrich W. Ebner-Priemer

Karlsruhe Institute of Technology (KIT)

Germany

Corresponding Author address:

Timothy J. Trull

210 McAlester Hall

Department of Psychological Sciences

University of Missouri-Columbia

Columbia, MO 65211

TrullT@missouri.edu

Note: The first author (TJT) is co-founder of TigerAware, a software platform used in designing and conducting ambulatory assessment studies. In addition, TJT is currently a Scientific Advisor for Boehringer Ingelheim Pharma.

IN PRESS, Journal of Abnormal Psychology

Abstract

The use of *ambulatory assessment* (AA; Trull & Ebner-Priemer, 2013) in psychopathology research, which includes experience-sampling methods (ESM) as well as ecological momentary assessment (EMA), has increased dramatically over the last several decades. Previously, methodological and reporting guidelines have been presented to outline best practices and provide input on methodological issues and decisions that are faced when planning and conducting AA studies (e.g., Bolger & Laurenceau, 2013; Mehl & Conner, 2012; Stone & Shiffman, 2002). However, despite the publication of these important resources and guidelines, it remains an open question as to how much uniformity or consistency is evident in the design and reporting of AA studies of psychopathology. To address this, we review the reported practices of published studies using AA in major psychopathology journals (*Journal of Abnormal Psychology, Psychological Medicine, Clinical Psychological Science*) over the last 7 years (2012-2018). Our review highlights: (1) sample selection and size; (2) sampling design; (3) selection and reporting of measures; (4) devices used and software; (5) compliance; (6) participant training, monitoring and remuneration; and (7) data management and analysis. We conclude with recommendations for reporting the features of future AA studies in psychopathology.

Key words:

ambulatory assessment (AA), reporting guidelines, ecological momentary assessment (EMA)

General Scientific Summary (GSS):

In clinical psychology, the use of daily-life research methods is increasing. However, the reporting of these methods is quite variable. We integrate existing best-practice guidelines, review articles from three major journals published over the last 7 years, report percentage of studies that meet these reporting guidelines, and offer recommendations for reporting the features of future AA studies in psychopathology.

The use of *ambulatory assessment* (AA; Trull & Ebner-Priemer, 2013) in psychopathology research has increased dramatically over the last decade. *AA*, which includes *experience-sampling methods (ESM)* as well as *ecological momentary assessment* (EMA; Stone & Shiffman, 1994), differs from traditional forms of assessment (e.g., self-report questionnaires, laboratory tasks, clinical and diagnostic interviews) in several important ways (Trull & Ebner-Priemer, 2013). First, because AA involves multiple assessments over time, it is uniquely suited to focus on *within-individual processes*. For example, depression is a dynamic process that may ebb and flow over time, often alongside contextual or environmental factors. Yet, traditional cross-sectional assessment requires individuals to somehow characterize their symptoms by aggregating in some unspecified way over extended periods of time (e.g., two weeks). Furthermore, traditional clinical assessment often requires some degree of retrospection (in extreme cases, over one's lifetime). In contrast, AA can be used to target momentary experiences (e.g., "within the last 15 minutes"), *minimizing retrospective biases and reliance on memory heuristics* (e.g., the peak-end rule; Fredrickson & Kahneman, 1993). AA captures slices of these processes in real- or near-real time, allowing an evaluation of not only mood processes (e.g., how much one's depression changes within and across days) but also potential internal and external influences on these processes. Thus, AA adds a needed *time dimension* to the assessment of psychological constructs. Finally, due to the collection of data during individuals' daily lives, the *ecological and external validity* of these assessments, by definition, exceeds that of more traditional measures that are completed in the artificial environment of the clinic, laboratory, or hospital.

AA is particularly well-suited to conduct research in psychopathology (Myins-Germeys et al., 2018; Trull & Ebner-Priemer, 2013). First, because AA integrates time into the assessment protocol, it can describe and characterize problematic mood states, mood changes, mood dynamics, and mood instability. Because emotion dysregulation involves, at least in part, mood changes and instability, and emotion dysregulation is recognized as a trans-diagnostic feature of psychopathology, AA has a wide

range of application in psychopathology research. Second, AA can also be used to assess and characterize problematic behaviors that are associated with psychopathology (e.g., substance use, binge and purge episodes, interpersonal conflict, non-suicidal self-injury, etc.). Third, AA can be used to assess problematic cognitions (e.g., rumination), expectancies (e.g., rejection sensitivity), and urges (e.g., craving, self-harm, etc.). Finally, because it incorporates reports of events, context, and individual differences, AA studies can evaluate the viability of proposed mechanisms that are associated with various forms of psychopathology (e.g., does increased negative affect precede substance use onset, and does substance use precede reports of lower levels of negative affect?).

Previously, reporting guidelines have outlined best practices and provided input on methodological issues and decisions that are faced when planning and conducting EMA and AA studies (e.g., Bolger, Davis, & Rafaeli, 2003; Bolger & Laurenceau, 2013; Fisher & To, 2012; Mehl & Conner, 2012; Stone & Shiffman, 2002). However, despite the publication of these and other important resources and guidelines, it remains an open question as to how much uniformity or consistency is evident in the design and reporting of AA studies in the field of psychopathology. Consistent reporting of methodological and analytical details of studies increases transparency, facilitates replication, and serves to enhance the rigor and utility of future studies (Kazak, 2018).

To address this, we integrated and organized recommendations from various guidelines into these modules: (1) sample selection and size; (2) sampling design; (3) selection and reporting of measures; (4) devices used and software; (5) compliance; (6) participant training, monitoring and remuneration; and (7) data management and analysis. Next, we reviewed all published studies using AA in three major psychopathology journals (*Journal of Abnormal Psychology*, *Psychological Medicine*, *Clinical Psychological Science*) published from 2012 to 2018. Specifically, we conducted a comprehensive, systematic literature search using Google Scholar to query for studies. Search terms combined ("experience sampling methods" OR "ecological momentary assessment" OR "ambulatory

assessment") with the name of each of the three specific journals (e.g. *source*: Clinical *source*: Psychological *source*: Science), and results were limited to studies published from 2012-2018 (inclusive). The initial search occurred in July 2018, updated with a search in November 2018.[1]

We reviewed titles and abstracts and evaluated articles returned from the searches for inclusion criteria. We included empirical studies that used some form of assessment at least at the daily level outside of a laboratory setting. See Figure 1 in the *Supplemental Materials* for PRISMA Flow Diagram of study selection and exclusion process. All articles were evaluated according to the reporting practices we and others (e.g., Fisher & To, 2012; Stone & Shiffman, 2002) recommend for AA articles (summarized in Table 1 and the text below). Approximately 76% (n=48) of the articles were also coded by a second person, and discrepancies were resolved by the authors.

We discuss the nature and importance of these reporting practices and present the results of our review in terms of the percentage of studies that met these reporting criteria. A listing of these proposed reporting practices in the text below are summarized in Table 1 as are the results from our review, organized by journal. We conclude with recommendations for reporting practices for future AA studies in psychopathology.[2]

Sample selection and size. Although not a guideline unique to AA research, it remains incumbent on investigators to provide a rationale and justification for the sample(s) being used as well as the number of participants required for analyses to possess the statistical power necessary to detect expected effects. First, participants that are included in AA studies should reflect the features and

[1] We chose these three journals for our review because (1) each is considered a top-tier outlet for psychopathology research; (2) a number of AA/EMA articles have been published in these journals over the last six years; and (3) readers of this journal, psychopathologists, are likely to access and publish in these outlets. However, we do not claim that this necessarily resulted in a representative sample of all articles using AA during this time period. Rather, we suggest that our review of these journals provides a best-case-scenario of reporting practices in the field of psychopathology given their impact and stature in psychopathology research.

[2] Of the 53 articles we identified in our review, only 3 of them reported the use of a non-self-report AA method of data collection. Therefore, our review focuses on self-report and does not discuss issues related to non-self-report AA methods. However, excellent recent reviews of issues associated with these methods exist and are presented (along with other resources) in the *Supplemental Materials*.

characteristics of the population to which the findings will be generalized. The sample should not be chosen primarily for convenience (e.g., undergraduates in a psychology class); rather, the sample should be suitable for drawing meaningful conclusions regarding the psychological theories or mechanisms relevant to the outcomes and processes of interest.

Second, as has been noted previously (e.g., Bolger & Laurenceau, 2013; Fisher & To, 2012), sample size selection should be based on a priori statistical power analyses or, if not available, investigators should demonstrate that the study is sufficiently powered for the effects of interest. At a minimum, to estimate power when planning a study, investigators must provide estimates of the number of participants in the study, the number of assessments completed by each participants (taking into account the likely average compliance rate), the anticipated effect size of interest, and intraclass correlation coefficient or ratio of between cluster variance to total variance (Arend & Schafer, 2019; Bolger & Laurenceau, 2013). Suggesting that sample size is sufficient because a previous study with the same sample size found significant effects is inadequate and potentially misleading. Fortunately, there are now a number of resources available to conducting power and analyses for multi-level models as well as for consulting tables in which major parameters are varied and power can be estimated based on simulations (e.g., Arend & Schafer, 2019; Bolger & Laurenceau, 2013; Lane & Hennes, 2019; see *Supplemental Materials*). Although we suspected that power analyses for AA studies would rarely be presented due to their complexity, we were somewhat shocked by our literature review. *In our review, only 2% of the articles explicitly reported that a power analysis was conducted either before data collection or after data analysis (Table 1).*

Finally, when the sample selection strategy places constraints on the range and types of responses made by participants, this should be noted in the Discussion section. For example, a study of inpatients may uncover different emotional and contextual triggers than a study of individuals in their natural environments.

Sampling design. Next, crucial to AA studies is the selection of and rationale for the sampling schedule (Bolger & Laurenceau, 2013; Mehl & Conner, 2012; Fisher & To, 2012; Stone & Shiffman, 2002). At one end of the spectrum is a simple end-of-day (EOD) assessment in which participants rate their aggregated experience for that day. For example, a participant might be asked about their overall mood, stress level, or number of times an event (e.g., interpersonal conflict) or behavior (e.g., consumed alcohol) occurred. Although convenient, this design may not be well-suited to test many theories of interest. AA methods are particularly useful in investigating momentary experiences and processes. Therefore, many AA researchers seek to collect data from different time-points each day. In these cases, sampling may be **random** throughout the day (most appropriate when the construct of interest is believed to be dynamic and fluctuating; e.g., mood), **interval-based** (to assess times or intervals that are meaningful for the construct of interest; e.g., EOD if meaningful, or 6pm to capture the end of a typical work day), or **event-based** (the participant initiates an assessment when a pre-defined event occurs; e.g., finishing an alcoholic drink). Some of the most interesting and powerful designs combine these types of assessments (random, interval, event-based) to provide a very rich, dynamic picture of how processes may unfold in daily life. For example, by combining assessment schedules, a researcher may uncover mood before, during, and after substance use.

There are important considerations that may impact the sampling schedule. For example, how many random (or interval-based) assessments are necessary to capture the expected variability of the construct (e.g., the temporal dynamics of negative affect) but minimize burden on the participant? This, of course will also depend on decisions regarding the length of the study (in days). In making these decisions, especially, we observe some neglect in estimating the base rates of events. If the major event of interest (e.g., opioid use) has a low base-rate in the sample that is being assessed, then a much longer study (with fewer unnecessary random prompts) must be planned to capture enough events within individuals. Another consideration concerns the dynamics and processes related to events that occur in

episodes (e.g., drinking episode), for which the intentional scheduling of prompted follow-up assessments can be extremely useful. Here, a researcher might choose to administer follow-up assessments 30 min, 60 min, and 90 minutes after the predetermined event or behavior to achieve *high density sampling* (Bolger & Laurenceau, 2013; Fisher & To, 2012; Stone & Shiffman, 2002). Furthermore, to capture episodes of varying lengths, an investigator might choose to "reset" the follow-up schedule if additional events are reported in any of the follow-ups. Finally, investigators should report the time frame of assessments (e.g., 9am to 9pm), and they should justify why only sampling certain hours of the day or night is appropriate for the research question. For example, in studying substance use, it is likely that the evening assessment time frame should be expanded (e.g., to midnight) and allow for event-based assessments to be initiated 24 hours per day.

To increase compliance and reduce burden, participants should be provided with a method for suspending prompts in advance. For example, when participants know they will be temporarily unavailable to answer prompts (e.g., while attending a movie or place of worship; while driving), they can use a "suspend" button provided by the device software to specify a suspension window (e.g., for the next 60 minutes) that will prevent and cancel any scheduled prompts during that time.

Finally, the technical details of sampling (e.g., prompting and recording practices; procedures for event-based entries; ability to suspend or delay responses; details on branching, triggering assessments, follow-ups or dense sampling of events/experiences) should be reported in the Methods section (Fisher & To, 2012; Stone & Shiffman, 2002). In addition, descriptive statistics like the mean (SD) time between prompts as well as the mean time elapsed between the lab and field phases of AA studies that combine lab-based and AA assessments are important to include in the Methods section.

Based on our review, only 17% of studies provided a rationale for their adopted sampling design, only 17% discussed their choices for sampling density (e.g., assessments per day) and scheduling (i.e., when the assessments are scheduled), and only 32% provided technical details of their studies' sampling.

Selection and reporting of measures. Previous guidelines (e.g., Fisher & To, 2012; Shrout & Lane, 2012; Stone & Shiffman, 2002) emphasize the importance of reporting psychometric properties of items and scales used in AA studies. Unfortunately, relatively few self-report measures have been validated across AA studies and samples; instead, researchers often select items from a larger cross-sectional measure and adapt the instructions to fit the desired timeframe (e.g., "over the last 15 minutes"). Despite the temptation of ease and convenience, we cannot assume that cross-sectional measures will retain original, or even similar, psychometric properties when administered repeatedly over shorter intervals. For example, basic descriptive measures like means and standard errors sometimes differ dramatically for mood ratings across differing time frames (e.g., right now, last 2 hours, last 24 hours, last week, etc.; Walentynowicz et al. 2018).[3] *In our review, 78% of the articles provided specific information on the content of the items administered, but only 30% of papers reported the psychometric properties of their chosen items (i.e., multi-level reliability; validity).*

This relative neglect of psychometric evaluation of AA questionnaires may be due to a lack of familiarity with methods to assess psychometric properties of repeated longitudinal data. Fortunately, there are options for addressing this issue (e.g., Calamia, 2019; Cranford et al., 2006; Fisher & To, 2012; Geldhof, Preacher, & Zyphur, 2014; Shrout & Lane, 2013). Notably, current publication standards require the evaluation and reporting of psychometric properties of scales used in traditional, cross-sectional studies, but have yet to require the same for measures and scales used in AA studies (e.g., Appelbaum et al., 2018).

It appears that the reliability and validity of AA measures are assumed, but not evaluated nor reported. For purposes of reliability and validity within the AA framework, it is recommended that complex constructs be assessed with at least three items, while discrete phenomena or behavior may be

[3] Given that different reporting timeframes require participants to access different sources of emotional knowledge which in turn have differing levels of cognitive demand (Robinson & Clore, 2002), investigators should consider issues of cognitive impairment in participants when deciding on rating timeframes.

assessed with a single item (Shrout & Lane, 2011). Furthermore, procedures for evaluating the psychometric properties of scales used in AA research have been outlined, in terms of reliability both within- (i.e., across time) and between-subjects (Geldhof et al., 2014; Shrout & Lane, 2011), as well as reliability of change scores within person (Cranford et al., 2006).

Devices and software used. Crucial to reports of AA studies are detailed descriptions of the devices (e.g., smartphones or external devices and sensors) and software used by participants, as well as how items are presented (e.g., scaling; response options), any branching or triggered follow-up assessments, and how sensor data (from smartphones or external devices) are collected. Concerning smartphones or electronic diaries, hardware and software versions should be reported to promote replicability (Fisher & To, 2012; Stone & Shiffman, 2002). *In our review, 76% of papers reported on the hardware (e.g., PDAs, smartphones) and software used in the study, although this was often a quite general description (e.g., "Android smartphone").*

There are mixed recommendations regarding whether participants can or should use their own smartphones in AA studies. On the one hand, participants would rather carry only one device and are likely more adept at navigating their own smartphone. However, the software adopted for the study may not operate as designed on some phones due to hardware/software incompatibilities, and/or participants may be prone to ignore study prompts while they are using their own smartphone for regular tasks. For these reasons, some investigators assign study-dedicated smartphones for their studies. Notably, a recent meta-analysis of EMA studies focusing on substance use indicated that the compliance rates for those that used their own phone in comparison to those that used study-provided phones did not differ significantly (Jones et al., 2018).[4]

[4] It is worth noting that studies that require participants to own a compatible smartphone will necessarily exclude those who do not have and possibly cannot afford these devices. Therefore, it is ideal for investigators to provide devices in these circumstances. Furthermore, apps should have the ability to collect data offline so that those without a data plan can still participate.

Finally, for those AA studies that incorporate internal (smartphone) and external (wireless) sensor data collection, it is important to note sampling frequency (i.e., how often are data collected or updated), how artifacts were identified and data were cleaned, and details about malfunctions (how sensor malfunctions were defined, frequency and systematic patterns of malfunctioning, e.g., late at night). Several excellent reviews and guidelines the use of device-based and wireless sensors appear in the *Supplemental Materials*.

Compliance. Given that AA research seeks to provide a window into the real lives of participants, non-response to surveys (or failure to wear sensors) on the part of participants will challenge the ability to generalize findings to one's typical daily life experiences. Therefore, it is crucial that AA investigators: (1) define compliance and lack of compliance (e.g., does a cancelled assessment due to suspension count as non-compliance?); (2) not only report overall compliance, but also compliance for each type of assessment (e.g., morning report, random reports, follow-up assessments; bedtime or EOD reports); (3) report both the mean level of compliance for each type of report as well as the range of compliance across participants; (4) describe and justify the thresholds for compliance necessary for participants to be included in the analyses (e.g., 75%); although there are no hard and fast rules for a specific threshold, it is important to note that as more missing data are included it may be harder to assume data are missing at random and the estimates for lagged effects become less reliable; (5) compare groups of participants for compliance rates; (6) and examine the data for systematic influences or patterns on compliance rates (e.g., time of day; day of week; day of study). *In our review, 65% of papers defined, generally, what constituted compliance (or missing data) and presented descriptive statistics on overall compliance to prompted reports.*

Participant training, monitoring, and remuneration. Compliance is substantially enhanced if participants are trained in study procedures, use of the smartphones/software, and proper wearing of external devices or sensors. Many studies do not mention whether or how much participant training

occurred prior to the start of AA data collection. Some exemplary examples of training and monitoring might include in-person meetings with investigators at the beginning of the study and as needed during the study, "practice" survey administration and completion while in the investigators' lab, and daily participant monitoring of compliance with intervention via phone or email if necessary. Concerning the remuneration schedules, many investigators use incentives to enhance compliance such as regular (e.g., weekly) in person meeting to provide compliance stipends, pro-rating payments if compliance fall below certain thresholds (e.g., 80%), and providing extra incentives for high levels of compliance over extended periods of time (e.g., over 90% for the entire study). Investigators should report the remuneration schedule and amount, how and whether completion is rewarded in the moment, whether participants can view their progress in the study and upcoming surveys, and any reported reasons for low compliance, for example. *Our review indicated that 73% of studies described procedures used to enhance compliance in their participants.*

Data management and analysis. AA studies present many challenges for data management and for data analyses (Bolger & Laurenceau, 2013; Mehl & Conner, 2012; Shiffman, 2014). First, the sheer amount of data collected far exceeds that from most traditional cross-sectional studies in that each participant may contribute large numbers of assessments depending on the sampling scheme (i.e., number of assessments per day) as well as the length of the study. For example, a study of 100 participants that provide an average of six assessments per day for 21 days would produce approximately 12,600 lines of data (which in turn may include 50+ variables per line of data!). Because of this volume, it is critical that data are collected, structured, and cleaned with great care and with an eye toward the ultimate analyses that will be performed. For example, once the data are cleaned, at a minimum each line of data should: (1) be associated with an ID number indicating which participant contributed the data; (2) have time- and date-stamps (indicating the start and finish time for each assessment) and be sorted; (3) include each prompt (answered or not) with a code for the type of

prompt or survey (e.g., morning report, random, event-based); (4) contain the notation for important study features that may serve as covariates in the analyses (e.g., day of study, day of week, weekend vs. weekday, etc.); and (5) document suspension with time-stamps (where applicable). Furthermore, each column (variable) should have a unique identifier (e.g., "distress" rated during random prompts) to avoid confusion or extra programming at later stages of data analyses. This may result in very long lines of data for each assessment, with missing values for many variables; however, this will make analyses more straightforward.

One under-appreciated step is to examine the distributions of variables and, for dependent variables, to assess the partitioning of variance into within- or between-person variability (e.g., see Bolger & Laurenceau, 2013; Shrout & Lane, 2011). If there is little variance attributable to within-person levels, the investigator may consider constraining analyses to examine only between-person effects. Fortunately, for most variables of interest, there is enough variability within-person to proceed with traditional multilevel models. In addition to reporting the intraclass correlations for major variables of interest (i.e., partitioning the variance into within- and between-person variability), we also recommend reporting the means and standard deviations (both within person and between person) in an Appendix.

The analytic models for major analyses should be clearly specified, ideally in formulas (at least in an Appendix), and the analyses should map onto the hypotheses of interest. It is also important to specify the statistical software (and version) used as well as the specific analytic modules (or options) used within the software package. Across software packages, defaults used in the specific analyses may vary, affecting the values of the estimates or effects. Reporting statistical software, packages, and programs used is important given that many researchers "ride the defaults". Researchers should specify and justify modeling decisions including: centering at different levels of analysis, modeling random versus fixed effects (slopes), aggregation of variables (at any level), and assumptions regarding distributions of the variables. Ideally, the software code/script used will be provided in an Appendix. In

addition, researchers must be clear about which covariates are included in the model, as well as any interactions tested (and, at what level of analysis). In this way, other investigators will have a clear idea how the data were structured, assumptions made about the data, and how the data were analyzed (and by which software packages and modules). These practices will help replication efforts and serve as instructional guidelines for decisions that the data analyst faces. *In our review, 60% studies provided detail on preparation of the data for analyses (including centering decisions), and 90% provide sufficient detail and rationale for the actual data analyses and models used.*

Discussion

Imagine your colleague or student wanted to design an AA study, for example, on the momentary (within-subject) effect of rumination on psychological distress. She would face several central questions, including: 1) Which items reliably capture the phenomena of interest (e.g., rumination, distress)? 2) What is the appropriate time-based design (i.e., how quickly does rumination affect distress; what is an optimal sampling time interval; how many days must individuals be monitored to observe enough instances of rumination and distress)? 3) Will the number of items and the proposed sampling time interval and length of the study place too much burden on participants, affecting compliance? 4) How many participants and assessments are necessary to evaluate the assumed association?

Our review suggests that answers to these four questions may be hard to come by. For example, psychometric properties of AA items, calculated within a multilevel context, were reported in only 30% of the studies. Compliance to AA protocols was not always presented, and multilevel power analyses were only presented in a handful of the articles. Not only was a basic description of methods and procedures sometimes missing, but, also, it was rare to find discussion on the **rationale** for these methodological decisions. Reporting that 10 assessments per day were used is beneficial, but an explanation that 10 assessments were chosen because pilot studies revealed better compliance with 10

assessments compared to 12 or 14 is much more informative. Similarly, a low sampling frequency (e.g., every 5 hours) may fail to "carve out" the time period necessary to observe lagged relationships reliably, whereas a high sampling frequency (e.g. every hour) might reveal these. We encourage researchers to provide not only a detailed description of their methodology but also a compelling rationale for their decisions as well as limitations and consequences of these decisions.

Of course, AA is an emerging methodological field, one that has only recently become more mainstream in the study of psychopathology, and, for instance, examples and software code for power analyses that consider multiple levels of analysis have been widely disseminated only recently. However, the accumulation of both methodological and substantive knowledge across AA studies would be greatly facilitated by better reporting of crucial features of studies, as well as explicit rationales for choices made by researchers designing their own studies. Whereas ambulatory assessment in the past was performed by a few expert groups, there are now many new users; this is a great development and testament to the power and attractiveness of AA to address important issues in psychopathological research. However, at the same time, this makes the explicit reporting and discussion of methodological aspects of AA studies even more critical.

In addition, the systematic evaluation and aggregation of findings from AA studies using meta-analyses fundamentally depends on clear and explicit reporting of study characteristics. In a recent meta-analysis of AA studies on substance use and addiction, Jones et al (2018) reported only 33% of the reviewed studies provided information on either excluding participants due to inadequate compliance or the number of participants who did not achieve minimum requirements for responding. The generalizability of meta-analyses on AA methodology is limited if reports on the methodological aspects of studies is infrequent or unclear. In short, the robustness of AA findings concerning psychopathology features and processes will only be possible if researchers provide in-depth descriptions of the

methodological features of their studies as well as the rationale for the decisions made in designing their studies.

However, we do not want to come across as too pessimistic. Although few articles included in our review were positively evaluated across all criteria in Table 1 (including our own!), we did note exemplary examples of meeting each criterion of reporting. This suggests to us that researchers are motivated to provide adequate reports of and rationales for their methodology; however, the lack of reporting may be due to a lack of consensus for reporting standards and, in addition, journals' space constraints. In the latter case, we encourage authors to take advantage of the option of submitting supplementary material and appendices when permitted in order to provide comprehensive reports of the design features of their studies.

Papers published more recently do appear to report more details regarding methodology, which is encouraging. However, brief reports, multi-method papers (e.g., combining fMRI and AA data), and secondary papers using data sets from previously published studies often do not adequately describe even the most basic aspects of their AA methodology. In these cases, a full report of the methodology can be made in an Appendix. Online supplements might also provide a full report of all items used in the AA study with information about their origin, time frame and instructions, and psychometric properties, for example. Toward this end, in the *Supplemental Materials*, we provide a template for reporting our recommended criteria that can be used by researchers. We see our recommended reporting guidelines template as a living document that should be updated and revised in the future.

More in-depth description of the statistical models and formulas, with software code, would be desirable as well. In several papers we could only guess whether psychometric properties were calculated, appropriately, in a multilevel framework or if data points were collapsed within participants. In summary, more detailed reports of methodological aspects (and rationales) will increase the informative value of our papers, which should be our main interest as researchers.

Finally, given that one explanation for unclear adherence to best practices and to reporting guidelines may be due to a lack of exposure to excellent resources, we compiled a selected list of resources that address the features of AA studies we highlighted in this review, as well as aspects of AA studies that we did not discuss in this review (i.e., guidelines for studying special populations of psychopathology; guidelines for analyzing intensive longitudinal data; guidelines for the use of wireless sensors, mobile phone sensors, cognitive or behavioral tasks, and collection of biological samples; and participant training and enhancing compliance). These resources appear in the *Supplemental Materials*, and, like the reporting template, our intention is that that this list will be a living document that can be updated to meet the needs of current and future researchers using AA to study psychopathology.

References

aan het Rot, M., Hogenelst, K., & Schoevers, R. A. (2012). Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical psychology review*, *32*(6), 510-523.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3-25.

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, *24*(1), 1-19.

Baltasar-Tello, I., Miguélez-Fernández, C., Peñuelas-Calvo, I., & Carballo, J. J. (2018). Ecological momentary assessment and mood disorders in children and adolescents: a systematic review. *Current Psychiatry Reports*, *20*(8), 66.

Bell, I. H., Lim, M. H., Rossell, S. L., & Thomas, N. (2017). Ecological momentary assessment and intervention in the treatment of psychotic disorders: a systematic review. *Psychiatric Services*, *68*(11), 1172-1181.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*(1), 579-616.

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods*. New York, NY: Guilford.

Bolger, N., Stadler, G., & Laurenceau, J.-P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285-301). New York, NY, US: The Guilford Press.Burke, L. E., Shiffman, S., Music, E., Styn, M. A., Kriska, A., Smailagic, A., ... & Mancino, J. (2017). Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. *Journal of medical Internet research*, *19*(3), e77.

Bussmann, J. B., Ebner-Priemer, U. W., & Fahrenberg, J. (2009). Ambulatory activity monitoring: Progress in measurement of activity, posture, and specific motion patterns in daily life. *European Psychologist*, *14*(2), 142-152.

Calamia, M. (2019). Practical considerations for evaluating reliability in ambulatory assessment studies. *Psychological Assessment*, *31*(3), 285-291.

Carpenter, R. W., Wycoff, A. M., & Trull, T. J. (2016). Ambulatory assessment: New adventures in characterizing dynamic processes. *Assessment*, *23*(4), 414-424.

Coravos, A., Khozin, S., & Mandl, K. D. (2019). Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ digital medicine*, *2*(1), 14.

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin, 32*, 917-929.

Dias, D., & Paulo Silva Cunha, J. (2018). Wearable health devices—vital sign monitoring, systems and technologies. *Sensors*, *18*(8), 2414.

Engel, S. G., Crosby, R. D., Thomas, G., Bond, D., Lavender, J. M., Mason, T., ... & Wonderlich, S. A. (2016). Ecological momentary assessment in eating disorder and obesity research: a review of the recent literature. *Current psychiatry reports*, *18*(4), 37.

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior*, *33*(7), 865-877.

Foster, K. T., & Beltz, A. M. (2018). Advancing statistical analysis of ambulatory assessment data in the study of addictive behavior: A primer on three person-oriented techniques. *Addictive behaviors*, *83*, 25-34.Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*(1), 45-55.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological methods*, *19*(1), 72-91.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838-854.

Harari, G. M., Müller, S. R., Aung, M. S., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, *18*, 83-90.

Heron, K. E., Everhart, R. S., McHale, S. M., & Smyth, J. M. (2017). Using mobile-technology-based ecological momentary assessment (EMA) methods with youth: A systematic review and recommendations. *Journal of Pediatric Psychology*, *42*(10), 1087-1107.

Hofmans, J., De Clercq, B., Kuppens, P., Verbeke, L., & Widiger, T. A. (in press). Testing the structure and process of personality using ambulatory assessment data: An overview of within-person and person-specific techniques. *Psychological Assessment*.

Holmlund, T. B., Foltz, P. W., Cohen, A. S., Johansen, H. D., Sigurdsen, R., Fugelli, P., ... & Elvevåg, B. (2019). Moving psychological assessment out of the controlled laboratory setting: Practical challenges. *Psychological assessment*, *31*(3), 292-303.

Houben, M., Ceulemans, E., & Kuppens, P. (in press). Modeling intensive longitudinal data. In A. G. C. Wright & M. N. Hallquist (Eds.), *The handbook of research methods in clinical psychology*. New York, NY: Cambridge University Press.

Houtveen, J. H., & de Geus, E. J. (2009). Noninvasive psychophysiological ambulatory recordings: Study design and data analysis strategies. *European Psychologist*, *14*(2), 132-141.

Jacobson, N. C., Chow, S. M., & Newman, M. G. (2019). The Differential Time-Varying Effect Model (DTVEM): A tool for diagnosing and modeling time lags in intensive longitudinal data. *Behavior research methods*, *51*(1), 295-315.

Jones, A., Remmerswaal, D., Verveer, I., Franken, I.H., Robinson, E., Franken, I. H. A., Wen, C.K., & Field, M. (2018). Compliance with Ecological Momentary Assessment Protocols in Substance Users: a Meta-Analysis. *Addiction, 114,* 609-619.

Kazak, A. E. (2019). Editorial: Journal article reporting standards. *American Psychologist, 73*, 1-2.

Kjærgaard, M. B., Bhattacharya, S., Blunck, H. & Nurmi, P. in Proceedings of the 9th international conference on Mobile systems, applications, and services 307-320 (ACM, Bethesda, Maryland, USA, 2011).

Kleiman, E. (2017, September 24). Understanding and analyzing multilevel data from real-time monitoring studies: An easily- accessible tutorial using R. https://doi.org/10.31234/osf.io/xf2pw

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, *35*(1), 7-31.

Lane, S. P., & Hennes, E. P. (2019). Conducting sensitivity analyses to identify and buffer power vulnerabilities in studies examining substance use over time. *Addictive behaviors*, *94*, 117-123.

Lansdorp, B., Ramsay, W., Hamidand, R., & Strenk, E. (2019). Wearable Enzymatic Alcohol Biosensor. *Sensors*, *19*(10), 2380.

Liao, Y., Skelton, K., Dunton, G., & Bruening, M. (2016). A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE checklist for reporting EMA studies (CREMAS). *Journal of medical Internet research*, *18*(6), e151.

Mehl, M. R., & Conner, T. S. (2012). *Handbook of research methods for studying daily life.* New York, NY: Guilford Press.

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry*, *17*(2), 123-132.

Nalepa, G. J., Kutt, K., & Bobek, S. (2019). Mobile platform for affective context-aware systems. *Future Generation Computer Systems*, *92*, 490-503.

Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A. E. G., ... & Dunn, G. (2011). Experience sampling research in individuals with mental illness: reflections and guidance. *Acta Psychiatrica Scandinavica*, *123*(1), 12-20.

Piasecki, T. M. (2019). Assessment of alcohol use in the natural environment. *Alcoholism: clinical and experimental research*, *43*(4), 564-577.

Raugh, I. M., Chapman, H. C., Bartolomeo, L. A., Gonzalez, C., & Strauss, G. P. (2019). A comprehensive review of psychophysiological applications for ecological momentary assessment in psychiatric populations. *Psychological assessment*, *31*(3), 304-317.

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological assessment*, *31*(2), 226-235.

Rodriguez-Blanco, L., Carballo, J. J., & Baca-Garcia, E. (2018). Use of ecological momentary assessment (EMA) in non-suicidal self-injury (NSSI): A systematic review. *Psychiatry research*, *263*, 212-219.

Schlotz, W. (2019). Investigating associations between momentary stress and cortisol in daily life: What have we learned so far? *Psychoneuroendocrinology*, 105, 105-116.

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science*, *24*(2), 154-160.

Shiffman, S. (2013). Conceptualizing analyses of ecological momentary assessment data. *Nicotine & Tobacco Research*, *16*(Suppl_2), S76-S87.

Shiffman, S. (2014). Conceptualizing analyses of ecological momentary assessment data. *Nicotine & Tobacco Research, 16*, S76-S87.

Shrout, P. E., & Lane, S. P. (2011). Psychometrics. In M. R. Mehl, & T. A. Conner (Eds.), *Handbook of studying daily life* (pp. 302-320). New York, NY: Guilford.

Singh, N. B., & Björling, E. A. (2019). A review of EMA assessment period reporting for mood variables in substance use research: Expanding existing EMA guidelines. *Addictive behaviors*, *94*, 133-146.

Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., ... & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017-1034.Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine, 16*(3), 199-202.

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, *24*(3), 236-243.

Stone, A., Shiffman, S., Atienza, A., & Nebeling, L. (2007). *The science of real-time data capture: Self-reports in health research*. Oxford University Press.

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., & Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PloS one*, *9*(4), e95978.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151-176.

Wac, K., & Tsiourti, C. (2014). Ambulatory assessment of affect: Survey of sensor systems for monitoring of autonomic nervous systems activation in emotion. *IEEE Transactions on Affective Computing*, *5*(3), 251-272.

Walentynowicz, M., Schneider, S., & Stone, A. A. (2018). The effects of time frames on self-report. *PloS one*, *13*(8), e0201655.

Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data*. Oxford University Press.

Walz, L. C., Nauta, M. H., & aan het Rot, M. (2014). Experience sampling and ecological momentary assessment for studying the daily lives of patients with anxiety disorders: A systematic review. *Journal of anxiety disorders*, *28*(8), 925-937.

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. *Journal of medical Internet research*, *19*(4), e132.

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life. *European Journal of Psychological Assessment*, *23*(4), 258-267.

Wright, A. G., & Zimmermann, J. (in press). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological assessment*.

Table 1. Recommended reporting guidelines for AA studies and results from literature review[1]

Recommended Reporting Criterion	Journal of Abnormal Psychology *(n=37[2])*	Clinical Psychological Sciences *(n=16)*	Psychological Medicine *(n=10)*	Total *(n=63)*
Justify sample size (e.g., using a multilevel power analysis)	3%	0%	0%	2%
Explain rationale for the sampling design (e.g., random, event-based, etc.)	19%	19%	10%	17%
Explain rationale for sampling density (e.g., assessments per day) and scheduling (i.e., when the assessments are scheduled)	16%	25%	10%	17%
Provide technical details of sampling (e.g., prompting and recording practices; procedures for event-based entries; ability to suspend/delay responses; branching details, triggering assessments, follow-ups or dense sampling of events/experiences)	38%	25%	20%	32%
Report full text of items, rating timeframes, response options/scaling	78%	75%	80%	78%
Report psychometric properties of items in the current EMA-study (between and within person), as well as the origin of the items	38%	31%	0%	30%
Fully describe hardware and software used	73%	75%	90%	76%
Define valid and missing data (for participants broadly, and specific to individual EMA reports); report descriptive analyses regarding valid data (e.g., mean per person, range, % participants above and below 80% threshold)	65%	62%	70%	65%
Describe the procedures used to enhance compliance and participation (e.g., remuneration schedule, participant training)	78%	69%	60%	73%
Describe the final data set: number of reports (total; person average; group average), days in study and retention rates, and rates of delayed or suspended responding (if applicable)	49%	50%	30%	46%
Preparation for data analyses: describe centering of predictor variables and at what level; report covariates included in the models	68%	50%	50%	60%
Data analysis: Describe levels of analysis (momentary, day, person); explain how time is taken into account in analyses; specify and justify choices of random versus fixed effects in models; describe analytic modeling used as well as statistical software used	95%	81%	90%	90%

Note. [1] The list of all articles reviewed appear in the Supplemental Materials.

[2] *N*=number of articles represented within each journal. Percentages indicate the percentage of articles that met this criterion.