Registered Report

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
July-September 2026, Vol. 9, No. 3, pp. 1–24

DOI: 10.1177/25152459261453933
www.psychologicalscience.org/AMPPS
Sage

Mapping Methodological Variation in Experience-Sampling Research From Design to Data Analysis: A Systematic Review

Lisa Peeters[1], Wim Van Den Noortgate[1,2], Marie Annelise Blanchard[1,3], Marie Suenaert[1], Gudrun Eisele[4], Olivia J. Kirtley[4], Richard Artner[1], and Ginette Lafit[1]

[1]Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium; [2]ITEC – An IMEC Research Group at KU Leuven, Leuven, Belgium; [3]Psychological Sciences Research Institute, Université Catholique de Louvain, Louvain-la-Neuve, Belgium; and [4]Center for Contextual Psychiatry, Department of Neurosciences, KU Leuven, Leuven, Belgium

Abstract

The experience-sampling method (ESM) has become a widespread tool to study time-varying constructs across many subfields of psychological and psychiatric research. Variety in subfields of research and constructs of interest has contributed to considerable methodological variation. Therefore, in this systematic review, we aimed to (a) describe the methodological variation in ESM-study designs and (b) assess the transparency (i.e., reporting and open-science practices) of ESM studies. We developed an extensive list of data-extraction items covering the entire workflow of an ESM study, from conception of the research question to reporting the results. These data were extracted from 150 recently published articles describing 162 studies applying ESM in the field of psychology and psychiatry. As expected, variation was observed for every methodological decision. Some findings aligned with previous syntheses of the literature (e.g., a median study duration of 10 days, a majority of studies including convenience samples of the general population, the high prevalence of multilevel modeling). Regarding other decisions, this review was able to provide new insights (e.g., questionnaire length and content often varied between and/or within participants, average sample sizes have increased over time). Transparency in reporting and open-science practices seem to have improved in recent years, but there is room for further development of awareness and guidance. This broad overview of ESM research serves as a first step toward a broader goal to improve the methodological quality of ESM research in psychology, contributing to a more rigorous, credible science of daily life.

Keywords

ESM, experience-sampling method, ecological momentary assessment, ambulatory assessment, study design, methodological variation, transparency, systematic review, open data, open materials, preregistration

Received 3/21/26; revision accepted 4/15/26

The experience-sampling method (ESM) has become the "gold standard" to study time-varying psychological processes, behaviors, and contexts in daily life. Examples include emotions (e.g., Blanke et al., 2020; Kalokerinos et al., 2019), substance use (e.g., Kurten et al., 2022; Weiss et al., 2022), and social experiences (e.g., Achterhof et al., 2022). The method is used across many subfields of psychological research, such as clinical psychology (e.g., Ader et al., 2022), personality psychology (e.g., Beck & Jackson, 2020; Kaurin et al., 2023), and organizational psychology (e.g., Shi et al., 2024). Also known by the

Corresponding Author:
Lisa Peeters, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium
Email: lisa.peeters@kuleuven.be

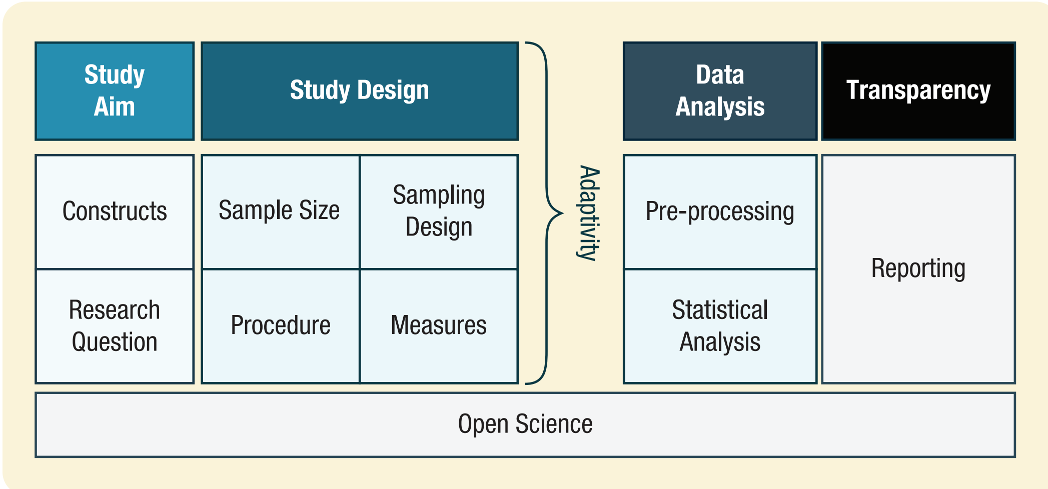

Fig. 1. The workflow of an experience-sampling-method study.

closely related terms "ecological momentary assessment" and "ambulatory assessment," ESM is an intensive, longitudinal data-collection method aiming to capture constructs of interest in real-time and real-life contexts (Myin-Germeys & Kuppens, 2022). In their daily life, participants are prompted, usually multiple times per day, to complete self-report measurements. These measurements can inquire about participants' momentary thoughts, feelings, behavior, and symptoms and the context in which the measurement takes place (e.g., location, company; Myin-Germeys et al., 2009). Through the integration of data collection in participants' daily life, ESM allows the collection of ecologically valid data and a reduction of recall biases (i.e., biases that occur because participants' ability to retrospectively report past experiences is often challenged; Beal, 2015). In the past, participants were typically asked to fill in a pen-and-paper questionnaire each time they received a prompt from a beeper (Csikszentmihalyi & Larson, 1987). Technological advancement has facilitated the application of ESM by allowing participants to receive prompts and fill out questionnaires on a smartphone or specialized research devices.

Despite its advantages and the technological advancement, ESM research does not come without its own challenges. Throughout the process of conducting an ESM study, the researcher is presented with a myriad of—often ESM-specific—choices and considerations, mapped in Figure 1. In the first part of this work, we give a broad overview of the workflow of an ESM study and present some of the literature that has aimed to map the methodological variation in (parts of) this process. This provides the backdrop for the second part of this work, a systematic review of the current ESM literature in psychology and psychiatry.

The Workflow of an ESM Study

The first decision researchers must make is whether an ESM study is appropriate for their research question. This is the case only if at least one of the constructs of interest varies over a relatively short time frame (Myin-Germeys & Kuppens, 2022), such as state body image (Stieger et al., 2022) or affective experiences (Achterhof et al., 2022).

Once it has been decided that ESM is appropriate for the research question and the constructs involved, the researcher must consider a number of complex design decisions. Choices regarding the sample size in an ESM study include the number of participants and the number of measurements per participant. The sampling design (e.g., time between measurements, type of sampling scheme) is closely related to the sample-size decisions in an ESM study. Similar to non-ESM research, a power analysis can be conducted for sample-size planning (Lakens, 2022). Unfortunately, because of the multilevel nature and the temporal structure of the data, conducting a power analysis can be complex (Lafit et al., 2021). A number of other design-related decisions regarding the measures (e.g., use of an existing validated questionnaire vs. creation of new items; Eisele et al., 2024) and the procedure (e.g., incentives, software/devices, briefing) must also be considered carefully (Fritz et al., 2023).

On one hand, design decisions must be based on the research question. For example, the number of assessments per day should follow the (theoretical) temporal dynamics of the construct(s) of interest (Seizer et al., 2024; Wright & Zimmermann, 2019). On the other hand, researchers must be mindful of the burden that design decisions place on participants and the quality (e.g., careless responding) and quantity of the resulting data. This is particularly relevant because methodological research has shown that there is a link between data quality and/or compliance and certain aspects of the study design, such as the type of sampling scheme (Himmelstein et al., 2019) and questionnaire length (Eisele et al., 2022).

In recent years, there has been a growing interest in adapting certain aspects of the design, such as the questionnaire length or the sampling scheme, to the participants, context, or incoming data (Schneider et al., 2024). "Adaptivity" is quite a broad term that as of yet is not used often in the ESM literature. Based on related literature in the field of educational sciences (Wauters et al., 2010), we formulated a definition of adaptivity in ESM designs. This definition is intentionally very broad so that the current work can maximally inform a more precise conceptualization of adaptivity. We define *adaptivity* as the adjustment of one or more characteristics of the ESM-study design to the individual participants' characteristics and preferences, preceding measurements, and/or context. This adaptivity can take place before the start of the data collection, in which case, it is referred to as "static adaptivity" (e.g., tailoring of the questionnaire items to individual characteristics; Scholten et al., 2022). "Dynamic adaptivity," on the other hand, refers to real-time updates to the study design during data collection (e.g., "episode-contingent" designs in which an event of interest triggers a follow-up period of more frequent measurements; Revol, Ariens, et al., 2023; Schreuder et al., 2024). Based on our definition, conditional questionnaire branching depending on context (e.g., depending on the reported social context; Achterhof et al., 2022) is considered dynamic adaptivity. Adaptive designs can improve the data quality/quantity (Cai et al., 2022), reduce participant burden (Bos et al., 2022), and/or improve parameter estimates (Revol, Ariens, et al., 2023), but researchers should be aware that they also bring forth additional complexities across the course of a study.

After the data collection, the raw data have to be processed (e.g., calculation of scale scores, missing-data handling), which can be guided by several existing guidelines and recommendations for preprocessing (e.g., Trull & Ebner-Priemer, 2020; Viechtbauer, 2021b). Revol, Carlier, et al. (2023) created an extensive framework to assist ESM researchers with preprocessing their data, including a website, templates, and an R package. The importance of careful consideration of preprocessing steps has been emphasized by recent findings that different preprocessing choices can lead to differences in the statistical results (Lafit et al., 2024; Weermeijer et al., 2022). After preprocessing the data, the researcher makes a number of decisions regarding data analysis (e.g., type of statistical analysis, software). For statistical analysis, many resources are available (e.g., Bringmann et al., 2013; Lafit, 2022; Viechtbauer, 2021a). Because of the hierarchical nature of the data (i.e., measurements nested within participants), resources almost always include an introduction to multilevel modeling (e.g., Bolger & Laurenceau, 2013). In addition, the temporal nature of the data (i.e., repeated measurements close together in time) often necessitates a statistical analysis that takes a possible autocorrelation between measurements into account. Common methods to capture this temporal dependency are autoregressive and vector-autoregressive (VAR) models (Ariens et al., 2020), which can be extended hierarchically (Bringmann et al., 2013).

Finally, researchers report their research to the scientific community. There are many ways in which meticulous reporting contributes to scientific progress. Thorough reporting of the entire research process allows reviewers and readers to make their own judgment about the quality of a study (Schiavone et al., 2023; Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020; Van Roekel et al., 2019). It also allows other researchers to replicate or reproduce the results, which is essential to the credibility of scientific findings (Artner et al., 2021; Simons, 2014), especially considering the "replication crisis" in psychology (Wiggins & Christopherson, 2019). Furthermore, complete reporting of the research process allows the reader to interpret the results in a more informed manner, considering the fact that a different choice at any point in the "garden of forking paths" (Gelman & Loken, 2013) could have led to different results. In addition to high-quality reporting, open-science practices can also increase the credibility of research findings in psychology. Examples of such open-science practices are registration, open materials, data sharing, and code sharing. Preregistration (i.e., recording the research plan before data collection) and postregistration (i.e., recording the analysis plan after data collection but before data access) aim to prevent exploitation of the garden of forking paths (Benning et al., 2019). Sharing study materials, for example, by contributing ESM items to the ESM Item Repository (Kirtley et al., 2024) increases measurement transparency and facilitates replication. Data and code sharing further increase transparency and allow for reproduction of the data analysis by independent researchers. Fortunately, many guidelines are available for reporting (Flake & Fried, 2020; Schiavone et al., 2023; Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020) and open science in ESM research (preregistration, Kirtley et al., 2021; data sharing, Alter & Gonzalez, 2018; sharing materials, Eisele et al., 2024; Kirtley et al., 2024).

All of the above can be summarized as follows: Across all stages of an ESM study, researchers face a large number of complex methodological decisions, each of which requires careful consideration.

Mapping the Methodological Variation

The methodological variation has been acknowledged by many researchers, but few have attempted to systematically assess the methodological choices made by ESM researchers. Such a systematic assessment is a complex endeavor because of the breadth of research questions that prompt researchers to conduct ESM studies. Wrzus and Neubauer (2023) found that emotions were a topic of study in the majority of ESM studies in the psychological literature, but other popular topics included a wide range of psychological concepts, such as personality states and health behaviors (Perski et al., 2022). There is also a wide variety in populations studied and sampling procedures in ESM research. Healthy (young) adult convenience samples have been found to be the most common, but ESM studies are also conducted with clinical or mixed samples and with different age groups and can be recruited in various ways (Wrzus & Neubauer, 2023). Existing systematic reviews on the use of ESM have often focused on specific populations (e.g., clinical populations) or specific topics of research. Examples are reviews of the use of ESM designs in adolescent samples (Van Roekel et al., 2019), psychosis research (Bogudzińska et al., 2024; Deakin et al., 2022), research on suicidal ideation (Ammerman & Law, 2022; J. J. Janssens et al., 2024), and pediatric health care (Van Dalen et al., 2023).

Furthermore, several methodological meta-analyses have assessed aspects of ESM-study design in the context of their relation to compliance, retention, or data quality. Again, most methodological meta-analyses have focused on a specific demographic or topic of research (e.g., substance use, Jones et al., 2019; severe mental disorders, Vachon et al., 2019). In their broad meta-analysis of compliance and dropout in ESM studies in psychology and related disciplines, Wrzus and Neubauer (2023) provided an overview of some important sample and design characteristics. They selected a subsample ($N = 477$) out of the 1,523 articles that fulfilled all inclusion criteria and coded ESM-design characteristics, sample characteristics, and compliance and dropout.

Other nonsystematic attempts to map methodological variation in ESM designs further attest to the broad variation in design decisions. K. A. M. Janssens et al. (2018) conducted a qualitative study with an international sample of 47 researchers who had experience conducting ESM or diary studies. These researchers reported their design choices in a recent study and the reasoning behind their decisions. Even in a sample of only 47 studies, there was a large variety in certain aspects of the study design (e.g., sampling frequency, number of items). Rationales for these decisions were categorized into four main themes: statistical reasons, feasibility, reliability, and nature of the variables. The most common rationales differed for the different design choices that were assessed. Reliability of the measurements was an important rationale for the chosen study duration, nature of the items (retrospective vs. momentary), the maximum response delay, and the sampling scheme (fixed vs. semirandom vs. random). Statistical reasons were also important for the choice of sampling scheme and study duration. The chosen measurement frequency was mostly motivated by participant burden (K. A. M. Janssens et al., 2018).

The common thread through this literature is the enormous breadth of the methodological variation in EMS-study designs. Regarding the sample, most of these studies coded the sample size and found a wide range (e.g., participants: M = 136.6, SD = 176.0, range = 4–2,001; Wrzus & Neubauer, 2023). The total number of measurements per participant was assessed in most cases by recording, on one hand, the study duration in days (e.g., Mdn = 17, range = 1–270; K. A. M. Janssens et al., 2018) and, on the other hand, the number of measurements per day (e.g., Mdn = 5, range = 1–50; K. A. M. Janssens et al., 2018). However, none of the aforementioned reviews or meta-analyses assessed whether a justification was provided for the reported sample size and/or number of measurements (e.g., a power analysis; Lakens, 2022).

Many methodological reviews and meta-analyses have also assessed certain aspects of the sampling scheme. The findings regarding the most common sampling schemes are mixed: K. A. M. Janssens et al. (2018) found that 52% of the included studies used a fixed sampling scheme, whereas only 18% did so according to Vachon et al. (2019). Signal-contingent sampling was found to be most common, but other types of sampling (e.g., event-based) are often reported as well (Wrzus & Neubauer, 2023). Only the qualitative study by K. A. M. Janssens et al. assessed researchers' justifications regarding the chosen sampling schemes; existing reviews and meta-analyses did not.

The scarcity of systematic assessment of ESM measurements is striking given that there is a rising interest in assessing the quality of measurements in the psychological ESM literature (e.g., Brose et al., 2020; Cloos et al., 2023; Horstmann & Ziegler, 2020). Many previous reviews have recorded only the number of items per questionnaire and have found, again, a lot of variation. For example, in their meta-analysis, Vachon et al. (2019) found that the number of items in their set of primary ESM studies varied between two and 135. Other aspects of the measurements that have been assessed in a systematic manner, albeit more rarely, are response scales and questionnaire development. Findings indicated that Likert scales are the most common response scales (Vachon et al., 2019) and that most studies use nonvalidated questionnaires (Deakin et al., 2022; Horstmann & Ziegler, 2020; Wright & Zimmermann, 2019).

Regarding the ESM procedure, it is evident that the use of electronic devices has become much more popular than pen-and-paper questionnaires (Ammerman & Law, 2022; Vachon et al., 2019). Incentives are often of a financial nature (i.e., direct monetary payment/gift cards; Wrzus & Neubauer, 2023) and are often compliance-based (Jones et al., 2019; Van Dalen et al., 2023). However, especially in health-care settings, it is increasingly common to receive personalized feedback as an incentive (Van Dalen et al., 2023). Despite the importance of (de)briefing for compliance and data quality (Palmier-Claus et al., 2011), very few studies have systematically assessed (de)briefing in ESM studies. However, it has been found that many ESM studies include at least some training before the ESM period (Deakin et al., 2022; Van Roekel et al., 2019).

The aforementioned literature has undoubtedly provided valuable insights into the process of designing an ESM study. However, several arguments can be made for a need for further investigation of the research process in ESM studies. First, the field of ESM has and continues to evolve quickly (Mestdagh & Dejonckheere, 2021). ESM designs continue to become more popular, more advanced, and often more complex (e.g., dyad studies, Griffith & Hankin, 2025; episode-contingent sampling designs, Revol, Ariens, et al., 2023). Therefore, these previous findings might not be representative of the field of ESM research today. Second, as mentioned previously, most previous reviews and meta-analyses have been population-specific or subject-specific. Although these works are of immense importance in their target subfield, one cannot assume that they are representative of the larger psychological ESM literature. Third, the methodological choices that were assessed in previous research were limited to a set of common design choices. Wrzus and Neubauer (2023) coded eight design features related to the sample size and sampling design (number of participants, total number of assessments, assessments per day, study duration) and the study procedure (incentives, reinforcement, participant care). Other reviews and meta-analyses coded similar sets of design choices. Although these choices remain relevant today, they are only a subset of the myriad of methodological choices researchers are faced with over the course of an ESM study. Out of all the steps of the workflow of a study (Fig. 1), the study design is the only one that has been reviewed extensively in the ESM literature. However, a study does not end after its design, and the quality of its final product (communication of the newfound knowledge to the scientific community) depends on much more than the quality of the design alone. ESM researchers again face complex choices during the data collection, when they are preprocessing/analyzing the data and when they report their work (Fig. 1).

Because of inconsistency in describing adaptivity, synthesis of the literature is difficult. As a consequence, to the best of our knowledge, no reviews or guidelines for adaptive methods in ESM research exist. As mentioned previously, guidelines and recommendations for preprocessing and analyzing data do exist. However, these resources provide only normative information. They do not describe the actual approaches taken by ESM researchers. In practice, preprocessing is often limited or not reported in detail (Blanchard et al., 2023; Revol, Carlier, et al., 2023). In a crowdsourcing project, Bastiaansen et al. (2020) asked 12 teams of researchers to use the same raw ESM data set to select target symptoms for treatment. The preprocessing included few steps for most teams, although all teams applied at least some form of preprocessing (e.g., clustering items, detrending, imputation of missing data; Bastiaansen et al., 2020). Researchers also rarely report checking statistical-model assumptions before conducting the analyses (Blanchard et al., 2023).

Concerning data analysis, Bastiaansen et al. (2020) and Blanchard et al. (2023) found that analyses are most often conducted using R (R Core Team, 2024) and that VAR models are the most common statistical models fitted to the data. Almost all teams in Bastiaansen et al. (2020) used a VAR model to investigate autoregressive effects (i.e., whether a variable predicts itself at a later time point) and cross-lagged effects (i.e., whether a variable predicts another variable at a later time point). Several teams additionally investigated contemporaneous effects (i.e., whether a variable covaries with another variable at the same time point). Some also used network approaches, in which multivariate data are analyzed by modeling statistical relationships between nodes in a network of variables (Borsboom et al., 2021). However, no two teams took exactly the same (or even very similar) approaches. Consequently, no two teams selected the same target symptoms for treatment. These results illustrate the critical role of the analysis strategy for the conclusions of a study. Moreover, they emphasize the importance of reporting clearly and completely how the results and conclusions of a study came to be.

However, current reporting practices do not necessarily reflect the abundance of resources available. In a review of ESM studies published in three major psychopathology journals between 2012 and 2018, Trull and Ebner-Priemer (2020) found a lot of room for improvement regarding reporting practices. For example, only 30% of the included studies reported psychometric properties of the items used in the ESM measurements, and only 32% reported the technical details of the sampling. Furthermore, only 17% provided a rationale for their sampling design, density, and scheduling. Regarding open-science initiatives as well, reviews have found that their application in practice is far from perfect. For example, many preregistered studies deviate from their preregistration without adequate disclosure of these deviations (Claesen et al., 2021). It is unclear how many researchers share their raw data and/or code.

The Current Study

We have illustrated the lack of a comprehensive overview of current practices in ESM research in psychology. The need for such an overview stems from two goals. On one hand, it allows for assessment and improvement of the quality of the use of ESM in psychological research. Knowledge of the current practices in ESM research can inform the development of guidelines and/or resources for all stages of the research workflow, which can, in turn, help researchers to apply ESM and report their work in an informed and methodologically valid manner. On the other hand, a broad overview of the current state of the art in ESM research can facilitate innovation and improvement of the method itself. One example is the use of adaptivity in ESM designs: We hope to map how ESM researchers apply adaptivity today and to use this information to identify how researchers can improve on these methods in the future.

To address this need, the aim of the current study is twofold: the description of the variation in current methodological practices in ESM research and the assessment of transparency in this field. "Methodological practices" here include the entire workflow of an ESM study, from conception of the research question to interpretation of the results (see Fig. 1). "Transparency" refers to both reporting practices and open-science practices.

To this end, a systematic review of the ESM literature was conducted. As discussed above, previous reviews were conducted in a specific subfield of psychology or a specific demographic and focused largely on a small set of design-related methodological choices. In contrast, in the current review, we aimed to describe a wide range of methodological choices in the entire field of psychological research. For each of these methodological choices, in this review, we assessed (a) whether they were reported clearly and completely; (b) the decisions made (as far as they were reported), including the use of overarching open-science practices (e.g., preregistration); and (c) the rationale given for some of the decisions. For this review, we developed an extensive list of items covering methodological choices across the entire process of conducting an ESM study, from initial conception to reporting of the results.

Method

Search strategy

The systematic review followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al., 2021). The key component of this search was the ESM in the context of psychological research. The following databases were searched: PsycARTICLES, MEDLINE, Psychology Databases (via ProQuest), Web of Science Core Collection (via Clarivate), and PubMed. To ensure the feasibility of the review, the search was limited to works published at most 1 calendar year before the start of the search, from January 25, 2024, until January 24, 2025. We argue that this focus on the recent literature reflects the aim of the review, that is, to assess the state of the art in psychological ESM research and to map the current research/reporting practices. Because the search was limited to 1 year, citation tracking would presumably identify very few additional records (if any) and was therefore not carried out. To ensure the full understanding of the text by all reviewers, the search was limited to records written in English.

Search terms were compiled with the help of experts in both systematic reviews and ESM. The search query included only terms related to ESM. No search terms related to psychology were included; whether the research belonged to the field of psychology was assessed in the screening phase. For the full search query, tailored to each of the databases, see the supplementary materials on OSF (https://osf.io/abvxp/files/3c7ta).

Screening

Records were eligible for inclusion if they fulfilled several criteria. They must describe an empirical study (i.e., no reviews or meta-analyses) in which ESM was applied with more than one measurement per day (i.e., no daily diary studies). Furthermore, ESM was used as a tool to answer a psychological-research question (e.g., whether momentary affect is related to social interactions; Achterhof et al., 2022), not as a topic of research itself (e.g., whether a certain aspect of an ESM design is related to compliance; Eisele et al., 2022). This criterion thus excluded feasibility and methodological studies. Only records that described the entire research process (i.e., no protocols) were included. Secondary data analyses were included if they reported all stages of the research process. Eligible records must be published in a journal that is classified as "psychology/psychiatry" in the Web of Science database. The "psychiatry" category was included to ensure that the review did not exclude a large portion of the clinical ESM literature, which is often published in psychiatric journals. For a full list of journals that fulfill this criterion, see the Supplementary materials on OSF (https://osf.io/abvxp/files/8gt2c). Records were excluded if passive-sensing data (e.g., accelerometry) were analyzed or if the study was not exclusively observational (e.g., intervention studies, experimental studies). Longitudinal studies that analyzed data from multiple ESM study waves (i.e., periods of ESM measurement separated by nonmeasurement periods; Myin-Germeys & Kuppens, 2022) were also excluded. Although these design features are of increasing importance in the ESM literature, they lead to a different, additional set of design and analysis choices that fall beyond the scope of the current review.

Study selection took place in two stages. In Stage 1, the titles and abstracts were screened. In Stage 2, the full texts of records that survived the first stage were scrutinized. In both stages, one reviewer evaluated all records, and several other reviewers independently screened a subsample to allow for the evaluation of consistency in decision-making. Reviewers used separate Zotero libraries and decision logs to ensure blinding of decisions. Because a skewed distribution of decisions (few inclusions and many exclusions) could lead to a low Cohen's kappa despite high values of percentage agreement (Belur et al., 2021), both statistics were calculated to assess interrater reliability. According to Landis and Koch (1977), kappa values between .61 and .80 can be interpreted as "substantial," and values above .80 are "almost perfect." Discrepancy in decisions was resolved through discussion. If the disagreement could not be resolved through discussion, an independent reviewer that was not involved in the screening stage resolved the issue. Discrepancies, discussions, and resolutions were recorded in a decision log[1]; full-text screening decisions were also recorded along with reasons for exclusions.

After all records were screened, a random sample of 150 records was selected out of all eligible records. Random sampling of eligible records has been used in systematic reviews (e.g., Wrzus & Neubauer, 2023) to allow for sufficient coverage of the literature while safeguarding the review's feasibility (considering the small team). The sample size was decided based on a pilot (described in the Supplemental Material) and a rough estimate of the number of relevant studies that are published in a single year based on previous systematic reviews (e.g., Wrzus & Neubauer, 2023). In case the number of eligible records would have been unexpectedly lower than this limit, we extended the search with an additional 6 months—to a total of 1.5 years before the start of the initial search.

Data extraction

The data extracted consist of main items (Fig. 2) with possible follow-up items in a branched structure (i.e., based on the response to the main item, different follow-up items are relevant). An initial set of items related to all stages of the research process was compiled based on a scoping search of older literature, available validity and reporting tools, and discussion with coauthors. The previously mentioned reporting recommendations (Flake & Fried, 2020; Schiavone et al., 2023; Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020) and meta-analyses (Wrzus & Neubauer, 2023) were particularly helpful in this endeavor. This set of items was subsequently refined in three rounds. In each round, four to six coauthors independently evaluated the list of items through Microsoft Forms. At the end of each round, a group discussion was held with all reviewers to improve the list of items. For details of the review process and the final list of data-extraction items including follow-up items, see the Supplemental Material. The last item, "Is the word 'Inertia' mentioned anywhere in the text?," was added in the interest of a separate study, and its results are not discussed in this work.

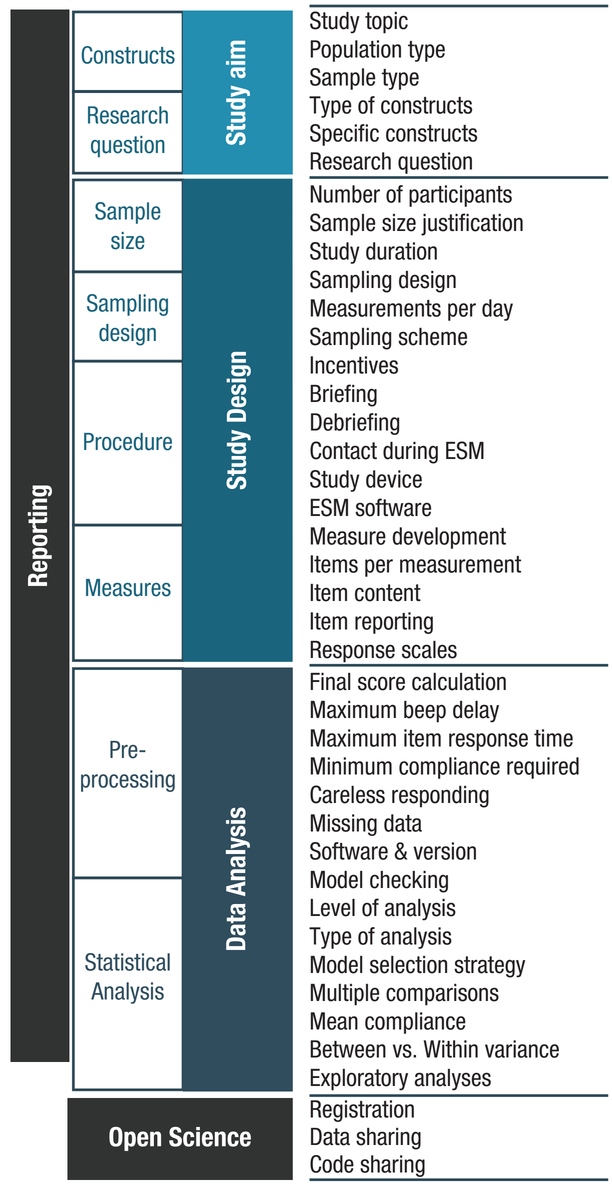

Fig. 2. Methodological choices covered by the data-extraction items.

For the data extraction, the reviewers made use of a data-extraction form created in Microsoft Forms, which supports the branched structure of the items. Again, one reviewer extracted data from all records, and several other reviewers extracted data from only a subsample to allow for the evaluation of consistency in decision-making. Missing data were coded as “not reported,” and no attempt was made to contact authors for more information. This is in line with the aim of this review because part of this aim is to assess the quality of scientific reporting in the literature. Thus, the nonreporting of any data is also considered valuable information.

If a record reported on more than one (sub)study that satisfied all inclusion criteria, data were extracted for each study separately. When studies reported on multiple research questions, up to three main research questions per study were considered. We included only research questions that were presented, including results, in the abstract of the record (following Artner et al., 2021). If more than three research questions satisfied this criterion, the most important research questions were selected through discussion with the research team, and precedence was given for research questions with a priori hypotheses.

The data-extraction form was piloted using records that were published in 2022 and would therefore not be included in the systematic review. For a full description of the pilot, see the Supplemental Material.

Data synthesis

For numerical items (e.g., the study duration in days), summary statistics were calculated. For binary or categorical items (e.g., whether a power analysis was conducted), proportions are reported. The synthesis of open-ended items is dependent on the item’s complexity: Simpler items (e.g., the statistical software that was used) were grouped into categories during data preprocessing, and proportions are reported. More complex items (e.g., a brief description of the statistical analysis) were synthesized in a narrative way. Because the synthesis of numerical and categorical data was largely predetermined, this synthesis was carried out by one researcher, and the reliability of the decisions was not assessed. For an RMarkdown document containing the preliminary analysis plan for these data, see https://osf.io/abvxp/files/nvtzf. Open-ended items were coded by two researchers independently, and reliability was assessed. Decisions regarding presentation of the numerical results and the narrative synthesis (Popay et al., 2006) of the open-ended items were discussed with multiple members of the research team. All statistical procedures were conducted using R (Version 4.4.3; R Core Team, 2024).

Results

Disclosures

This systematic review is the result of a registered report submitted to and peer reviewed by the Peer Community in Registered Reports. For the preregistration of the recommended Stage 1 manuscript, see https://osf.io/ztvn3. Small deviations from the Stage 1 protocol occurred. During the first round of screening, the wording of some exclusion criteria was not clear to all members of the team and was revised. However, only the wording in the screening manual changed, not the core ideas of the criteria. In addition, some items in the data-extraction form were changed to include a missing “not reported” option or to add extra clarifying instructions.[2] Finally, open items were coded by one researcher instead of the preregistered two because responses were either very straightforwardly numerical/categorical (e.g., the statistical software used) or too broad and/or subjective to be meaningfully coded into categories (e.g., justifications for certain sampling schemes). For the reviews and recommendation of the Stage 2 registered report, see https://rr.peercommunityin.org/articles/rec?id=1140.

The study materials, raw data, and analysis code used to produce the results are available on OSF (https://osf.io/abvxp/). The full data are also available in a dynamic web-based interface that allows users to easily navigate, filter, and group specific variables (https://airtable.com/appqKHsqcoFJqk9n3/shrqy53nz7EZDvn3y).

Study selection and interrater reliability

The search yielded 3,574 unique records, 473 of which passed the first stage of screening. In this stage, the percentage agreement between reviewers was 97.4%, and Cohen’s kappa was .887, signifying almost perfect interrater reliability (Landis & Koch, 1977). In the second stage of screening, 369 records fulfilled all inclusion criteria. The interrater reliability was lower than in the first stage (87.9% agreement, Cohen’s $\kappa = .694$) but still substantial (Landis & Koch, 1977). The final sample consists of 150 records (reporting on 162 studies), selected from all eligible records using a random number generator. The complete flowchart is shown in Figure 3.

For most categorical data-extraction items, interrater reliability was substantial ($\kappa > .60$) or almost perfect ($\kappa > .80$). For some items, kappa indicated less than substantial agreement because of a highly skewed decision distribution even though the percentage of agreement was high. However, a few items required further attention, and disagreements most often occurred on whether a decision was reported clearly enough. Furthermore, some concepts assessed in this review lack clear definitions (e.g., whether variation is adaptive), leading to

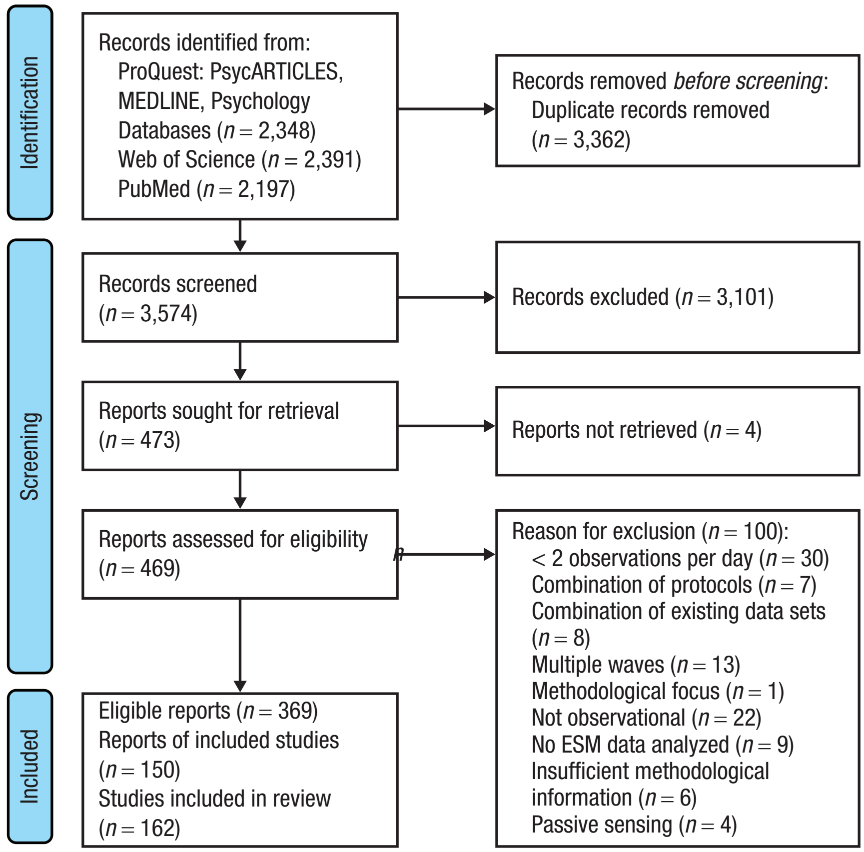

Fig. 3. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

different interpretations by different raters. In these cases, decisions were discussed within the team and if necessary, to ensure high-quality data, reassessed for all records. We must also note that although there was often overlap between raters for multiselect items, selections were often not exactly the same. Particular caution is warranted in the interpretation of the study topics and the type of research questions. For interrater reliability measures for all categorical and numerical items and a more in-depth discussion, see the supplementary materials on OSF (https://osf.io/abvxp/files/3c7ta).

Study aim

Information related to the study aim was reported in (almost) all studies. Only the clinical status of the sample was reported in fewer than 90% of all studies (Fig. 4). Table 1 presents descriptive statistics. Note that some categories have been collapsed for brevity; for a more extensive table, see OSF (https://osf.io/abvxp/files/6xyrp).

Person-level variables were measured and included in analyses in most studies (61.1%). Correspondingly, a majority of studies investigated relationships between time-varying and person-level variables (52.5%).[3] However, the most commonly examined relationships were lagged (56.8%) relationships between time-varying variables. Only 10 studies (6.2%) contained research questions about the relationship between data of two or more participants (e.g., affect dynamics in romantic couples). Almost all studies (89.5%) provided a justification for the use of ESM data to answer their research questions (Table 2).

The most common study topics were emotions (45.7%) and mental health (37.7%), followed by a substantial "other" category (25.9%), which included topics such as cognitive control, stigma, and moral behavior. The most commonly measured specific constructs, "negative affect" and "positive affect," were measured in 51 (31.5%) and 42 (25.9%) studies, respectively. Many other studies also measured emotions but conceptualized them more broadly (e.g., one-dimensional affect/mood) or more specifically (e.g., happiness, anger, shame). For all measured constructs see the full data at https://airtable.com/appqKHsqcoFJqk9n3/shrqy53nz7EZDvn3y/tblT1fQKdsNMcaVgk.

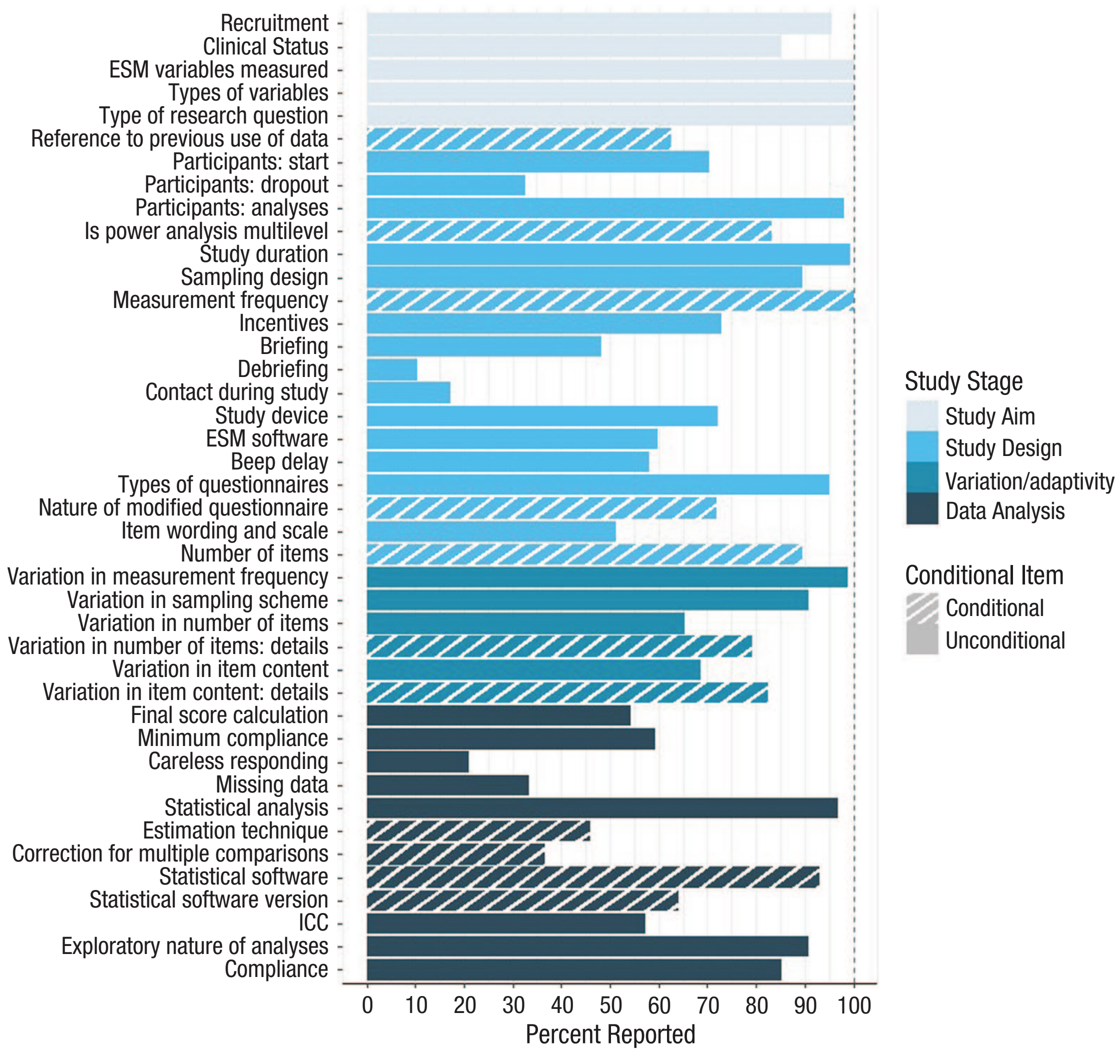

Fig. 4. Percentage of studies that reported each methodological decision.

Fifty-one studies (31.5%) reported including intentionally recruited clinical participants in their sample either with (n = 16) or without (n = 35) a nonclinical control group. However, the most common samples were general-population samples (47.5%) recruited through convenience sampling (67.9%).

First authors were primarily affiliated with universities in Europe (45.6%) and North America (42.6%) but also in parts of Asia (i.e., China, Taiwan, Japan, Singapore), Australia, and Egypt. For a figure depicting the location of all first-author affiliations, see supplementary materials on OSF (https://osf.io/abvxp/files/6xyrp).

Study design and adaptivity

As shown in Figure 4, almost all studies reported the number of participants included in the statistical analyses, the study duration, the measurement frequency, and the type of questionnaires used. Other decisions were reported less consistently. Some were even reported only sparsely (the number of participants that chose to discontinue their participation, briefing and debriefing, contact during the ESM period). Clear reporting about variation was more prevalent for the measurement frequency and sampling scheme than for the questionnaire length and content (Fig. 5).

A majority of studies (60.5%) used secondary data or data that were collected as part of a larger project. Despite this large percentage, there was little overlap in the data sets used in our sample, so we present results using all studies. Descriptive statistics of variables related to study design are available in Table 3.[4]

Incentives were most often financial (67.9%) and compliance-based (66.7%). A majority of studies used

Table 1. Descriptive Statistics of Variables Related to the Study Aim and Population (N = 162)

	n	Percentage
Type of variables		
Both time-varying variables measured with ESM and person-level variables	99	61.1
Only time-varying variables measured with ESM	63	38.9
Type of research question[a]		
Average level/fluctuations of one ESM variable	16	9.9
Concurrent relationships between ESM variables	73	45.1
Lagged relationships between ESM variables	92	56.8
Relationships between ESM and person-level variables	85	52.5
Other	14	8.6
Linked data of several participants		
No	152	93.8
Yes	10	6.2
Clinical status[b]		
General population (i.e., clinical participants not intentionally recruited but also not excluded)	78	48.1
Intentionally recruited clinical participants (with or without control group)	51	31.5
Nonclinical participants only (clinical participants are excluded)	10	6.2
Type of sample[b]		
Convenience sample	110	67.9
Online sample	9	5.6
Representative sample	6	3.7
Student sample	30	18.5
Study topic[a]		
Work/education	24	14.8
Emotions	74	45.7
Relationships (general, romantic, family)	37	22.8
Mental health	61	37.7
Personality states	7	4.3
Physical health and health behavior	24	14.8
Situations/places	2	1.2
Other	42	25.9

Note: ESM = experience-sampling method.
[a]Multiple options can be selected for each study.
[b]For information on the percentage of studies that reported each variable, see Figure 4.

smartphones for the ESM protocol (64.8%), and a wide range of ESM software was used. There was no obvious "most popular" app/website, although some were used more regularly (e.g., mEMA, https://ilumivu.com/; MetricWire, https://metricwire.com/; Wenjuanxing, www.wjx.cn). Almost half of all studies reported some kind of briefing on the protocol through informational materials (9.9%) or training by the researchers (40.1%).

Semirandom sampling designs were most common (46.9%), followed by fixed sampling designs (29.6%). However, justification for the choice of sampling design was rarely reported (22.1%; Table 2). The median study duration was 10 days with a wide range (4–90 days). The measurement frequency as well showed a wide range, from two to 21 observations per day (Mdn = 5). The measurement frequency was fixed for all participants in almost all studies with some exceptions, such as studies that measured more frequently during the weekends or studies that employed event-based sampling (Fig. 5). In the latter case, events might be more common on some days and for some participants, thus resulting in variation in the number of measurements per day. In contrast, a substantial number of studies varied the sampling scheme (i.e., the timing of the

Table 2. Number of Studies That Provided a Justification for a Selection of Design/Analysis Choices (N = 162)

	n	Percentage
Use of ESM	145	89.5
Sample size	78	48.1
Sampling type	32	22.1
Modification of existing questionnaire[a]	43	43
Adaptivity		
In measurement frequency[a]	5	26.3
In sampling scheme[a]	13	23.6
In questionnaire length[a]	16	28.1
In questionnaire content[a]	22	31.4
Statistical analysis	111	70.7

Note: ESM = experience-sampling method.
[a]Conditional item: the percentage refers to the percentage of studies that justified a particular choice compared with all studies that reported that choice.

measurements) between participants. In most cases, this adaptive variation took the form of adapting the sampling window to the schedule of the participants (e.g., work schedule, waking hours). For all forms of adaptivity, a minority of studies provided a justification (23.6%–31.4%; Table 2).

More than half of the studies (55.6%) reported a beep delay shorter than the interbeep interval (i.e., the time period in which participants were allowed to respond to questionnaires was shorter than the time between two scheduled beeps) and had a median of 30 min (range = 1.5–480). No studies reported a limit on the time participants could spend responding to each single item.

Existing questionnaires were used without modification in 37.7% of all studies, whereas almost two thirds of studies modified existing questionnaires (61.7%) or developed new items (64.2%). Eight studies did not provide sufficient detail about the origins of the questionnaires used. Validity and reliability information was reported more often for existing questionnaires[5] (37.9% and 63.8%, respectively) than for new questionnaires (10.6% and 26.9%, respectively). In fewer than half of the relevant cases was the modification of existing ESM or trait questionnaires clearly described (39.0%) or justified (44%; Table 2). The full wording and scales of all items were included in the text or supplementary materials in approximately half of all studies (50.6%).

There were relatively few studies in which the questionnaire length (28.4%) and content (21.6%) were reported to be fixed (Fig. 5). Common forms of questionnaire-related adaptivity were conditional item branching (i.e., between- and within-participants variation depending on the responses) and time-of-day-dependent questionnaires (i.e., within-participants variation). When the questionnaire length was reported to be fixed, the median number of items was 13.5 (range = 2–50). Although the Likert scale was most commonly used (74.4%), many studies used combinations with other item scales.

A median of 153 participants started the study, which dropped to 123 for the final analysis sample. In 10 studies, the sample size was justified using an a priori multilevel power analysis, and 64 studies provided a sample-size justification other than an a priori power analysis (e.g., recommendations in the literature, a post hoc power analysis). In total, 48.1% of studies provided some justification for their sample size (Table 2).

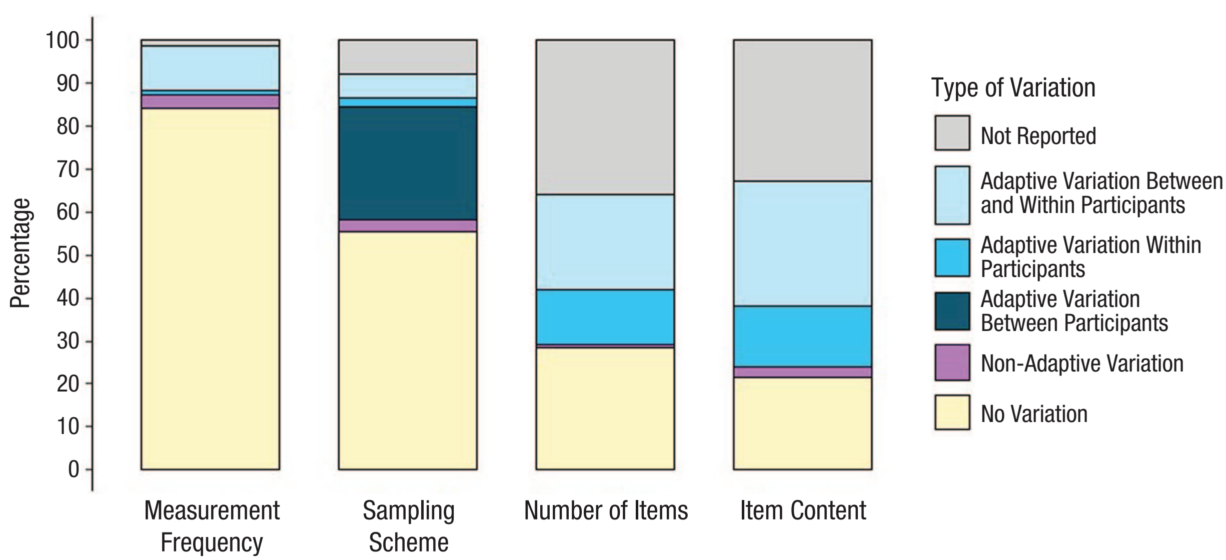

Fig. 5. Variation in measurement frequency, sampling schemes, questionnaire length, and questionnaire content.

Table 3. Descriptive Statistics of Variables Related to the Study Design (N = 162)

	n	Percentage
Incentives[a,b]		
Course credits	22	13.6
Feedback (immediate or later)	6	3.7
Financial (direct or lottery)	110	67.9
None	4	2.5
Compliance-based incentives (if financial or course credit, N = 117)	78	66.7
Briefing[a,b]		
Informative materials (e.g., letter, instruction video)	16	9.9
Briefing by researcher (in person or online)	65	40.1
Debriefing with discussion of ESM period[b]	6	3.7
Contact during ESM period[b]	22	13.6
Device[a,b]		
Smartphone	105	64.8
Other	14	8.6
Type of sampling design[a,b]		
Event-based sampling design	18	11.1
Fixed sampling design	48	29.6
Random sampling design	24	14.8
Semirandom sampling design	76	46.9
Other	4	2.5
Questionnaires[a,b]		
Existing questionnaire	58	35.8
Modified questionnaire	100	61.7
New questionnaire	104	64.2
Item scales[a,b]		
Binary/categorical response scale	75	46.3
Likert scale	121	74.7
Open item	28	17.3
Visual analogue scale	41	25.3
Other	10	6.2
A priori multilevel power analysis	10	6.2

	M	SD	Mdn (range)
Number of participants[b]			
Start	226.9	246.7	153 (1–1,382)
Dropout	20.2	44.7	4 (0–284)
Analyses	192.2	226.9	123 (1–1,913)
Study duration (days)[b]	13.2	10.1	10 (4–90)
Measurement frequency (per day; if fixed, N = 140)[b]	5.1	2.7	5 (2–21)
Questionnaire length (if fixed, N = 42)[b]	16.8	12.3	13.5 (2–50)
Beep delay (minutes)[b,c]	55.8	64.3	30 (1.5–480)

Note: ESM = experience-sampling method.
[a]Multiple options can be selected for each study.
[b]For information on the percentage of studies that reported each variable see Figure 4.
[c]Descriptives include only studies in which the beep delay was a fixed number of minutes. Four studies (2.5%) allowed participants to respond until the next beep. Nine studies (5.6%) reported varying beep-delay periods.

Data analysis and open science

As shown in Figure 4, few variables related to the data analysis—only the type of statistical analysis, the software used, and whether some analyses were exploratory—were reported in almost all studies. Assessment of careless responding, missing-data handling, model-estimation technique, and correction for multiple comparisons was reported in fewer than half of all studies.

Table 4 presents descriptive statistics[6] of variables related to data analysis. In 74 studies (45.7%), participants were reported to be excluded based on low

Table 4. Descriptive Statistics of Variables Related to the Data Analysis (N = 162)

	n	Percentage
Final score calculation[a,b]		
A summary of several items	117	72.2
One single-item score	110	67.9
Other	8	4.9
Exclusion based on low compliance[b]	74	45.7
Assessment of careless responding[b]	31	19.1
Missing-data handling[b]		
Imputation	7	4.3
Listwise deletion	18	11.1
Other	30	18.5
Statistical software[a,b]		
Mplus	48	29.6
R	96	59.3
IBM SPSS	22	13.6
Other	20	12.3
Some model assumptions are discussed	55	34.0
Type of variability modeled[a]		
Between and within subjects	148	91.4
Between subjects only	30	18.5
Within subjects only	4	2.5
Model estimation[a,b]		
Maximum likelihood	43	26.5
Bayesian	26	16
Other	6	3.7
Model selection[b]		
Theory driven	91	56.2
(Combination of theory and) data driven	70	43.2
Model selected and fitted on the same data (N = 70)	68	97.1
Multiple comparisons	93	57.4
Multiple comparisons correction[a,b] (N = 93)		
Benjamini-Hochberg	11	11.8
Bonferroni	6	6.5
Other	10	10.8
None	8	8.6
Exploratory analyses reported[b]	104	64.2

	M	SD	Mdn (range)
Minimum compliance for inclusion[b] (%; N = 72)	39.1	21.3	33.3 (2.9–80)
Average compliance[b] (%)	74.9	15.2	76.7 (21–97)

[a]Multiple options can be selected for each study.
[b]For information on the percentage of studies that reported each variable, see Figure 4.

compliance. In these studies, the median compliance requirement was one third of all beeps (range = 2.9%–80%). The median of the average compliance was 76.7% (range = 21%–97%). In the studies that reported missing-data handling, some commonly used methods were listwise deletion, full-information maximum likelihood, and the default settings for dynamic structural equation modeling (DSEM; Asparouhov et al., 2018) in Mplus (Muthén & Muthén, 2017). Other preprocessing steps, such as assessment of careless responding (19.1%) and model-assumption checks (34%), were not often reported to have been undertaken. Most studies used a combination of single-item scores and summary scores in their analyses.

Almost all studies reported analyses that modeled both between- and within-subjects variability, often with multilevel regression models (most commonly, linear regressions without an autoregressive effect) or multilevel structural-equation-modeling (SEM) approaches. However, around one fifth of studies (18.5%) conducted

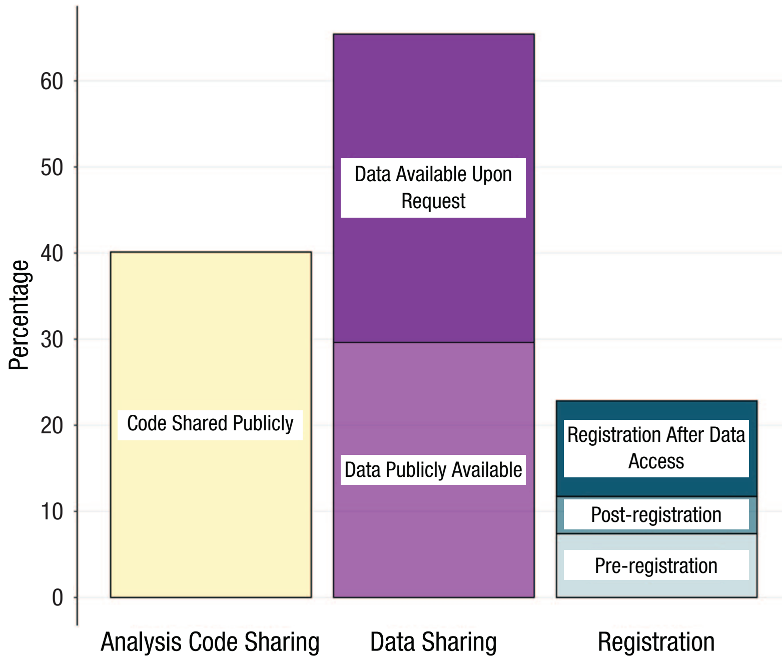

Fig. 6. Open-science practices.

analyses that focused on between-subjects variability using, for example, single-level linear regression or a comparison of summary statistics. Four studies (2.5%) focused on within-subjects variability. At least some justification for the choice of statistical analysis was provided in 70.7% of studies (Table 2). For model estimation, authors mainly reported using various maximum likelihood approaches and Bayesian estimation. Models were chosen based on theory in just over half of all studies (56.2%). In studies in which some data-driven model selection took place, the model was almost always (97.1%) fitted on exactly the same data that were used for the model selection without correction. Although multiple comparisons were common (57.4%), reporting a correction for multiple comparisons was not. When reported, the Benjamini-Hochberg correction was the most common. A majority of studies (64.8%) reported exploratory analyses. Statistical analyses were most often conducted in R (59.3%) and/or Mplus (29.6%).

For an overview of open-science practices, see Figure 6. Analysis code was reported to be publicly available in 40.7% of studies. Data are publicly available for 29.6% of studies and available on request in 35.8% of studies. Finally, some form of registration occurred in 22.8% of studies. However, only 7.4% of studies were preregistered (i.e., before data collection).

Discussion

This systematic review served two primary aims: assessing the transparency (i.e., reporting and open-science practices) of ESM studies and describing the methodological decisions that are made throughout the process of an ESM study. These two aims cannot be viewed separately because we cannot map methodological decisions that are not reported clearly. Note that we do not intend to discuss reporting practices in a normative manner because development of reporting guidelines is beyond the scope of this work (but see Limitations and Future Directions). Instead, we aim to provide reporting information in a descriptive manner to facilitate interpretation of methodological findings. Therefore, in what follows, we discuss these two aspects in tandem for each stage of the research process. We compare our findings with the existing literature and highlight new insights.

Study aim

Information regarding the study aim (e.g., research question, variables of interest) was reported clearly in almost all studies. Only the inclusion of clinical participants was reported slightly less often, mostly in nonclinical studies. One explanation for this could be that for studies that intentionally recruit clinical participants because this is relevant to their research question, it is evident that this information should be reported. When the research question of a study is not of a clinical nature and thus clinical participants are not intentionally recruited, researchers might deem it less relevant to report whether clinical participants were explicitly excluded.

As reported by previous syntheses (J. J. Janssens et al., 2024; Perski et al., 2022), almost all studies included a rationale for the use of ESM. To our knowledge, the types of research questions in ESM studies have not been systematically assessed before. We found that most studies included several types of research questions, often a combination of (lagged and/or concurrent) relationships between ESM variables, and relationships between those variables and one or more person-level variables. An example of this is a primary research question about the lagged relationship between momentary stress and momentary positive upregulating strategies with an additional investigation of the moderating effect of person-level depressive symptoms on this relationship (Colombo et al., 2024).

In general, our findings regarding study topics align with those of Wrzus and Neubauer (2023) in that emotions and mental health were the most common topics. However, using the same predefined categories, we found a large "other" category. This might point toward a broadening field of application of ESM. For example, there is an increasing interest in applying ESM in the field of cognitive psychology (e.g., Brearly et al., 2025; Doucette et al., 2024; Hakun et al., 2025).

The interest in emotions and mental health is reflected in the constructs that were measured. In line with other syntheses (Perski et al., 2022), an affect measure (e.g., positive/negative affect, specific emotions) was included in the vast majority of studies, including studies in which the research question was not primarily about affect. Many studies also measured constructs related to mental health (e.g., symptoms of affective disorders, substance use, disordered eating). Despite the sensitivity of the topic, 16 studies (9.9%) measured suicidal ideation/intent/behavior or nonsuicidal self-injury. This suggests that the burdensome nature of ESM does not prevent its use with even the most vulnerable populations in mental-health research, a field in which researchers are increasingly interested in studying within-persons processes (Myin-Germeys et al., 2018). Furthermore, approximately one third of the studies included intentionally recruited clinical participants (with or without a nonclinical control group), a number similar to previous general syntheses (Van Roekel et al., 2019; Wrzus & Neubauer, 2023).

It is unclear whether our findings regarding the relatively small number of studies that reported excluding participants based on clinical diagnoses is in line with previous findings because previous literature often reported the percentage of "healthy" samples (K. A. M. Janssens et al., 2018; Wrzus & Neubauer, 2023) without a separate "general population" category. We assume but cannot be certain that in those cases, "healthy" samples refer to both purely nonclinical samples and general-population samples (the latter of which are by far the most common in our review). The finding that convenience sampling is the most common recruitment strategy is in line with most of the literature (e.g., Deakin et al., 2022; Wrzus & Neubauer, 2023).

Study design and adaptivity

The sample size (with a median of 123 participants included in the analyses) was higher than reported in previous syntheses (e.g., $n = 90$ in Doucette et al., 2024; $n = 100$ in Perski et al., 2022; $n = 87.5$ in Wrzus & Neubauer, 2023).[7] This finding is quite surprising, taking into account that other syntheses did not explicitly differentiate between the sample size at the start of the study and the one included in the analyses, which could have inflated their estimates compared with ours (Heggestad et al., 2022). One possible explanation is that in the current review, we included only studies conducted very recently and awareness of the importance of sufficient sample sizes has increased in recent years (Vankelecom et al., 2025). This is also reflected in the finding that almost half of all studies included some justification for their sample size and that 11% conducted an a priori power analysis. Although these numbers may seem discouraging, they present a large improvement compared with the 2% and 0% of studies that provided sample-size justifications in the reviews by Trull and Ebner-Priemer (2020) and Van Roekel et al. (2019), respectively. Even compared with more recent reviews that reported between 6.6% and 8.3% of studies conducting a priori power analyses (Perski et al., 2022; Van Dalen et al., 2023), we see a positive trend.

As previously discussed, previous findings regarding sampling types are mixed. We found that semirandom sampling was the most common type (in line with Blanchard et al., 2023; Bogudzin´ska et al., 2024; Vachon et al., 2019) and that event-based sampling was relatively uncommon (in line with Ammerman & Law, 2022; Perski et al., 2022; Van Roekel et al., 2019; Wrzus & Neubauer, 2023). Fixed sampling—the prevalence of which previous literature disagreed the most on—was used in almost one third of all studies, more than reported in some literature (Vachon et al., 2019; Van Roekel et al., 2019) and less than reported in other literature (K. A. M. Janssens et al., 2018; Perski et al., 2022). Interpretation of these findings warrants two notes. First, the percentage of nonreporting of the sampling design was higher (10.5%) in our review than in all other literature discussed here without any obvious explanation.[8] Second, although we did not formally investigate this, we observed that some sampling types were more common in specific fields of research. For example, fixed sampling seemed to be more common in organizational research than in other fields. This might explain part of the discrepancy between the current review and previous literature that focused on a specific field of research (e.g., mental-health research; Vachon et al., 2019). Clear rationales for the chosen sampling type could explain why some types are more common than others (in specific fields or in general), but rationales were not often provided.

The study duration (*Mdn* = 10 days) and measurement frequency (*Mdn* = 5) observed in this review were within the range of other syntheses, although there is quite a bit of variability in the literature (e.g., Ammerman & Law, 2022; Deakin et al., 2022; Doucette et al., 2024; Van Dalen et al., 2023). Within-studies adaptivity in the measurement frequency was uncommon outside of the inherent variation that comes with event-based sampling, although some studies sampled more often on specific days of the week (e.g., weekends; Van Dalen et al., 2023). There was more variation in the distribution of the measurements throughout the day because many studies adapted the time window during which measurements occurred to the participants' personal schedules (e.g., sleeping times, work schedules).

Although explicit justifications for this adaptivity were not often provided, implicitly, we see two main reasons for the flexible time windows (and event-based measurements): (a) to safeguard compliance and lower burden (by ensuring participants are sampled when they are naturally awake/available) or (b) to measure only during periods when participants are doing an activity of interest (e.g., working, consuming alcohol). These reasons reflect what can be viewed as the main goals of adaptivity in the study design in general: improving participant experience and making sure measurements are maximally informative.

Similar goals can be assumed to underlie adaptivity in questionnaire length and content, which often (but not always) co-occur. Common types of questionnaire-related adaptivity observed in this review were branching based on previous responses (e.g., about the social context) and constructs being measured only at times during the day when they are most relevant (e.g., sleep in the morning, affect throughout the entire day, work-task completion at the end of the workday). These design features ensure that participants are not burdened unnecessarily by irrelevant questions and that constructs are measured only at informative times.

Partly because of this adaptivity and partly because of a lack of clear reporting (also noted by Van Dalen et al., 2023; Van Roekel et al., 2019), fewer than one third of all studies reported a fixed number of items. The questionnaire length in those studies (*Mdn* = 13.5) was comparable with some previous literature (Van Dalen et al., 2023) but lower than others (*Mdn* = 30, K. A. M. Janssens et al., 2018; *M* = 22.5, Vachon et al., 2019). It is possible that this can be explained by an increasing effort to avoid overburdening participants with long questionnaires in recent years (Eisele et al., 2022), although at this time, this is purely speculative.

Newly developed items and modified existing questionnaires were used in approximately two thirds of all studies, and unmodified existing questionnaires were used in just above one third. Most studies used a combination of questionnaire types. This aligns with the previous finding that two thirds of constructs in ESM research are measured using newly developed measures (Perski et al., 2022) and with the high prevalence of questionnaire modification reported by Heggestad et al. (2022). The latter reported shortening the questionnaire and adapting the time frame as the two most common types of modification, which is further supported by our data (in those cases that reported the modification clearly, which was a minority). Psychometric properties were generally only sparsely reported (in line with Deakin et al., 2022; J. J. Janssens et al, 2024), although reliability was reported for more than two thirds of the existing questionnaires. Although Heggestad et al. (2022) found that 83.3% of ESM studies reported reliability, they noted that the vast majority of the studies in their sample reported Cronbach's alpha, a less than ideal measure in the context of multilevel observations. Although we did not formally record specific reliability measures in our review, we can attest that many studies that discussed reliability reported multilevel measures (e.g., McDonald's omega), indicating an increased awareness of appropriate reliability measures for ESM data.

Our findings regarding the study procedure largely align with previous literature. Incentives were reported in approximately 70% of studies (comparable with Perski et al., 2022; Van Dalen et al., 2023, Wrzus & Neubauer, 2023) and were most often of a financial nature (Van Dalen et al., 2023; Wrzus & Neubauer, 2023) with a compliance-based incentive structure (which was slightly more common than observed in other reviews). As could be expected, the vast majority of studies used smartphones to collect ESM data. The evolution of smartphone use is clear when comparing these findings with reviews that include older studies, which often used personal digital assistants or even pen and paper (e.g., Bell et al., 2024; Deakin et al., 2022; Vachon et al., 2019). The specific software used was reported in approximately 60% of all studies. There was relatively little overlap in the commonly used apps/platforms in the current review and the two other reviews that previously reported this (Bell et al., 2024; Doucette et al., 2024), reflecting the broad variety of software available for ESM data collection. As noted in some previous literature (J. J. Janssens et al., 2024; Van Roekel et al., 2019), information regarding (de)briefing and contact during the ESM period was only sparsely reported. It is unclear whether this reflects a lack of reporting or an absence of (de)briefing/contact. Because not a single study explicitly reported not briefing participants on the study protocol, we can assume that the vast majority of studies did conduct some briefing, but half of the sample did not report this.

Data analysis and open-science practices

Preprocessing steps were reported only sparsely (20%–60%). Most studies used a combination of single-item scores and aggregate scores of multiple items in their analyses and used these scores on a momentary level. In addition, around one fourth of all studies included an aggregate score over all ESM measurements (often as a covariate). However, we must note that the calculation of the final scores used in the analyses was reported clearly in only half of the studies, and for the others, we deduced the final score-calculation method from other available information.[9] We saw a slight increase in studies reporting that they assessed careless responding compared with previous literature (Heggestad et al., 2022). Missing-data handling was, in most cases, reported only if this deviated from the default settings (although some studies explicitly reported using default Mplus settings). Considering the high prevalence of multilevel regressions conducted in R—for which the default option is listwise deletion of missing data—we can assume that the true percentage of studies that used listwise deletion is much larger than the reported percentage (10.5%; Blanchard et al., 2023). Around a third of studies reported some checks of model assumptions.

In line with Heggestad et al. (2022), approximately 60% of studies reported whether they excluded participants based on a minimum compliance cutoff. Almost all of those studies that reported doing so reported the exact cutoff (which ranged between 3% and 80%). Compliance was reported (or could be calculated based on available information) in most studies, and the average compliance of 74.9% was within the range reported in the literature (e.g., Deakin et al., 2022; Van Roekel et al., 2019; Wrzus et al., 2023). Note that our estimate includes both studies that reported compliance of the whole sample (before exclusions) and studies that reported compliance of only the final analysis sample (after participants with very low compliance had been excluded), so this should be interpreted with caution.

Almost all studies conducted at least one type of multilevel analysis (in line with Bell et al., 2024) and justified this based on the nested structure of the data. Previous investigations of data-analysis practices in ESM have focused explicitly on network analyses (Blanchard et al., 2023) or on a specific type of research question (treatment-target selection; Bastiaansen et al., 2020). This hinders comparison with findings of the current general synthesis, in which the most common type of analysis was (linear/logistic) multilevel regression without time-lagged predictors (as opposed to the very high prevalence of autoregressive models in the aforementioned literature). This may be surprising because lagged relationships between constructs were the most common type of research question. However, a significant number of studies incorporated this time lag in the study design (by measuring the predictor earlier in the day than the outcome) instead of in the data analysis. Others did model the time lag explicitly by using VAR-based models (Ariens et al., 2020), and there was a relatively high prevalence of DSEM (Asparouhov et al., 2018) compared with previous literature (Bastiaansen et al., 2020; Blanchard et al., 2023).

Just over half of the studies selected models based on theory alone, and the rest included at least some data-driven model selection (e.g., data-driven selection of covariates, fully data-driven machine-learning approaches). In the latter case, nearly all studies fitted their selected model on the same data that were used to select the model, leading to potential invalid postselection inference (Kuchibhotla et al., 2022). Specific estimation methods were sparsely reported, leading us to assume that most studies used the default estimation option for their analysis (software). In line with previous literature (Bastiaansen et al., 2020; Blanchard et al., 2023), a majority of studies used R to conduct analyses, although we observed an increase in the use of Mplus, possibly because of the increasing popularity of SEM-based methods.

A positive trend is apparent regarding open-science practices given that we found more registered studies, publicly available data, and publicly available analysis code than observed in previous reviews (Blanchard et al., 2023; J. J. Janssens et al., 2024). With registrations in almost a quarter, code shared in 40%, and data shared (publicly or on request) in more than 60% of studies, there is certainly room for further growth. Nevertheless, this positive trend is encouraging and shows the value of open-science resources and guidelines developed in recent years (e.g., Kirtley et al., 2021).

Limitations and future directions

In this systematic review, we succeeded in mapping methodological practices and transparency in psychological ESM research. Our findings build on previous literature but also provide new insights into methodological practices that were not previously investigated. However, this review is not without shortcomings. First, feasibility concerns limited the scope of the review and the size of the included sample. Intervention studies and studies that analyzed passive-sensing data were not included despite their increasing prevalence in the field (e.g., ecological momentary interventions; Bronas et al., 2024; Postma et al., 2024). Furthermore, fewer than half of the eligible records were included. However, because of the random study selection (following Wrzus and Neubauer, 2023), we are confident that our sample is an adequate representation of the recent[10] literature.

Second, some methodological choices proved to be difficult to code objectively and consistently despite extensive piloting of the data-extraction form. In some cases, discrepancies might have resulted from a combination of vague or inconsistent reporting and lack of consensus within the team on interpretations of the data-extraction items (e.g., adaptivity, recruitment). Through discussion of these cases within the team, we resolved these issues and were able to provide high-quality data. In other cases, particularly the more exploratory open-ended ones, items were simply not designed to be purely objective (e.g., What counts as a sufficient justification for a certain sampling scheme?). Here, some discrepancy was expected and did not hinder the narrative interpretations, but the subjectivity of these data should be acknowledged.

Third, the current synthesis of the data was purely descriptive and focused on each item individually regarding both the methodological practices and transparency. This was intentional, and we argued the value of such a comprehensive descriptive mapping of ESM research above. However, this is not all that can be learned from these data. Methodological choices interact: Researchers balance and consider a large web of decisions when conducting an ESM study. Further investigations of these data could replicate previously observed relationships (e.g., the trade-off between study duration and measurement frequency; Jones et al., 2019; Wrzus & Neubauer, 2023). Furthermore, the breadth of the methodological decisions recorded in these data provide a unique opportunity to further detangle the decision-making process of ESM researchers across all stages of a study (e.g., How do analytic choices align with the research questions asked and the temporal structure of the collected data?).

In addition, our findings regarding transparency invite further investigation. Some methodological decisions are reported more than others. However, information that was available only in supplementary materials was not counted as reported (with the exception of the full text and scales of questionnaire items), and authors can have good reasons to not report certain decisions in the main text. For example, some journals have strict word or page limits, forcing authors to carefully select which information is essential enough for the main text. In their reporting guidelines, Trull and Ebner-Priemer (2020) acknowledged this and recommended using supplementary materials to describe the full extent of the methodology. However, we note that information that is not reported in the main text may be missed by a casual reader who does not consult long supplementary materials.[11] The decision of which pieces of information are necessary to report in the main text can be subjective. Therefore, rigorous guidance should be developed in dialogue with ESM experts and applied researchers. The extensive list of data-extraction items created for this review could serve as a starting point for broad yet detailed reporting guidelines.

Conclusions

In this broad systematic review, we provided an overview of methodological decisions and transparency in psychological ESM research. Unsurprisingly, the broad range of applications of ESM is accompanied by a large variety in methodological practices. Further research will have to disentangle the relationships between individual methodological decisions and how they potentially differ between fields of research. Although there is a lot of room for growth regarding clear and complete reporting and open-science practices, comparison of our findings and reviews conducted in previous years inspires optimism. Detailed consensus-based guidelines could provide motivated researchers with the guidance necessary to further enhance the quality of ESM research and reporting in psychology.

Transparency

Action Editor: Felix J. Thoemmes
Editor: Felix J. Thoemmes
Author Contributions

Lisa Peeters: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Visualization; Writing – original draft; Writing – review & editing.

Wim Van Den Noortgate: Conceptualization; Methodology; Supervision; Writing – review & editing.

M. Annelise Blanchard: Investigation; Methodology; Writing – review & editing.

Marie Suenaert: Investigation; Writing – review & editing.

Gudrun V. Eisele: Conceptualization, Methodology, Writing – review & editing.

Olivia J. Kirtley: Conceptualization; Methodology; Writing – review & editing.

Richard Artner: Conceptualization; Methodology; Writing – review & editing.

Ginette Lafit: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Supervision; Writing – review & editing.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The research presented in this article was supported by research grants from the Research Council of KU Leuven (STG/23/050, C+/24/002, C16/23/011, C14/23/062) awarded to G. Lafit. M. A. Blanchard is a postdoctoral researcher of the Fund for Scientific Research - FNRS. G. V. Eisele was funded by the Research Foundation Flanders (1223725N). O. J. Kirtley was supported by research grants from the Research Council of KU Leuven (C+/24/009) and Research Foundation Flanders (G049023N, G1176725N).

Open Practices

This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Lisa Peeters https://orcid.org/0009-0001-0718-8160

Wim Van den Noortgate https://orcid.org/0000-0003-4011-219X

M. Annelise Blanchard https://orcid.org/0000-0002-9605-7022

Gudrun V. Eisele https://orcid.org/0000-0002-4466-3733

Olivia J. Kirtley https://orcid.org/0000-0001-5879-4120

Richard Artner https://orcid.org/0000-0002-4515-5650

Ginette Lafit https://orcid.org/0000-0002-8227-128X

Acknowledgments

We thank Enrico Guarino, who lent a helping hand in the screening phase of this review in the context of his research internship. We are also grateful to the two reviewers and the recommender at Peer Community in Registered Reports for providing helpful and constructive feedback during the peer-review process. The raw data, including full references for all records included in the review, are available at https://airtable.com/appqKHsqcoFJqk9n3/shrqy53nz7EZDvn3y or on OSF (https://osf.io/abvxp/files/5ey9a).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459261453933

Notes

1. The decision log is available on OSF (https://osf.io/abvxp/files/pju78).
2. The screening manual and data-extraction form are available in the supplementary materials on OSF (https://osf.io/abvxp/files/3c7ta); the deviations from the Stage 1 version are clearly indicated.
3. This percentage is lower than the percentage of studies that included person-level variables because some studies included person-level variables only as covariates and not as part of the main research questions.
4. Note that some categories have been collapsed for brevity, and a more extensive table is available on OSF (https://osf.io/abvxp/files/6xyrp).
5. This encompasses reporting of previous investigations of the questionnaire's psychometric properties (e.g., a validation study) and/or efforts to establish reliability/validity with respect to the study's own sample (e.g., calculation of reliability coefficients).
6. Note that some categories have been collapsed for brevity, and a more extensive table is available on OSF (https://osf.io/abvxp/files/6xyrp).
7. Note that for dyadic or triadic studies, each participant was counted separately. For example, a dyadic study with 50 dyads was coded as having 100 participants. However, even if each dyad was coded as one "participant," the median would decrease only slightly (to Mdn = 117).
8. We attempted to be very strict in our data extraction by not guessing or assuming things that were not reported explicitly, which might explain part of this observation. However, we did not see the same pattern for other data-extraction items for which we were similarly strict.
9. For eight studies, too little information was available to reasonably deduce how final scores were calculated.
10. The restriction to studies published in 1 year (January 2024 until January 2025) limits the generalizability to the entire psychological ESM literature, but this is a minor limitation considering this study's focus on the current methodological landscape of ESM research.
11. We thank one of the reviewers for this suggestion.

References

Achterhof, R., Kirtley, O. J., Schneider, M., Hagemann, N., Hermans, K. S. F. M., Hiekkaranta, A. P., Lecei, A., Lafit, G., & Myin-Germeys, I. (2022). Adolescents' real-time social and affective experiences of online and face-to-face interactions. *Computers in Human Behavior*, *129*, Article 107159. https://doi.org/10.1016/j.chb.2021.107159

Ader, L., Schick, A., Simons, C., Delespaul, P., Myin-Germeys, I., Vaessen, T., & Reininghaus, U. (2022). Positive affective recovery in daily life as a momentary mechanism across subclinical and clinical stages of mental disorder: Experience sampling study. *JMIR Mental Health*, *9*(11), Article e37394. https://doi.org/10.2196/37394

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *The American Psychologist*, *73*(2), Article 146. https://doi.org/10.1037/amp0000258

Ammerman, B. A., & Law, K. C. (2022). Using intensive time sampling methods to capture daily suicidal ideation: A systematic review. *Journal of Affective Disorders*, *299*, 108–117. https://doi.org/10.1016/j.jad.2021.10.121

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research*, *137*, Article 110191. https://doi.org/10.1016/j.jpsychores.2020.110191

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. https://doi.org/10.1037/met0000365

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., De Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., . . . Bringmann, L. F. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, Article 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Beal, D. J. (2015). ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(2), 383–407. https://doi.org/10.1146/annurev-orgpsych-032414-111335

Beck, E. D., & Jackson, J. J. (2020). Consistency and change in idiographic personality: A longitudinal ESM network study. *Journal of Personality and Social Psychology*, *118*(5), 1080–1100. https://doi.org/10.1037/pspp0000249

Bell, I. H., Eisner, E., Allan, S., Cartner, S., Torous, J., Bucci, S., & Thomas, N. (2024). Methodological characteristics and feasibility of ecological momentary assessment studies in psychosis: A systematic review and meta-analysis. *Schizophrenia Bulletin*, *50*(2), 238–265. https://doi.org/10.1093/schbul/sbad127

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research*, *50*(2), 837–865. https://doi.org/10.1177/0049124118799372

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

Blanchard, M., Contreras, A., Kalkan, R., & Heeren, A. (2023). Auditing the research practices and statistical analyses of the group-level temporal network approach to psychological constructs: A systematic scoping review. *Behavior Research Methods*, *55*(2), 767–787. https://doi.org/10.3758/s13428-022-01839-y

Blanke, E. S., Brose, A., Kalokerinos, E. K., Erbas, Y., Riediger, M., & Kuppens, P. (2020). Mix it to fix it: Emotion regulation variability in daily life. *Emotion*, *20*(3), 473–485. https://doi.org/10.1037/emo0000566

Bogudzińska, B., Jaworski, A., Zajdel, A., Skrzypek, K., & Misiak, B. (2024). The experience sampling methodology in psychosis risk states: A systematic review. *Journal of Psychiatric Research*, *175*, 34–41. https://doi.org/10.1016/j.jpsychires.2024.04.050

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. The Guilford Press.

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., Van Borkulo, C. D., Van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*, Article 58. https://doi.org/10.1038/s43586-021-00055-w

Bos, F. M., Von Klipstein, L., Emerencia, A. C., Veermans, E., Verhage, T., Snippe, E., Doornbos, B., Hadders-Prins, G., Wichers, M., & Riese, H. (2022). A web-based application for personalized ecological momentary assessment in psychiatric care: User-centered development of the PETRA application. *JMIR Mental Health*, *9*(8), Article e36430. https://doi.org/10.2196/36430

Brearly, T. W., Elbich, D. B., Dhinojwala, M., & Hakun, J. G. (2025). Contextualized cognition: Clarifying associations between remote unsupervised performance and clinically relevant contextual factors. *The Clinical Neuropsychologist*. Advance online publication. https://doi.org/10.1080/13854046.2025.2534720

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLOS One*, *8*(4), Article e60188. https://doi.org/10.1371/journal.pone.0060188

Bronas, U. G., Marquez, D. X., Fritschi, C., Petrarca, K., Kitsiou, S., Ajilore, O., & Tintle, N. (2024). Ecological momentary intervention to replace sedentary time with physical activity to improve executive function in midlife and older Latino adults: Pilot randomized controlled trial. *Journal of Medical Internet Research*, *26*, Article e55079. https://doi.org/10.2196/55079

Brose, A., Schmiedek, F., Gerstorf, D., & Voelkle, M. C. (2020). The measurement of within-person affect variation. *Emotion*, *20*(4), 677–699. https://doi.org/10.1037/emo0000583

Cai, L., Barnes, L. E., & Boukhechba, M. (2022). A framework for adaptive mobile ecological momentary assessments using reinforcement learning. In K. Arai (Ed.), *Intelligent systems and applications* (pp. 31–50). Springer International Publishing.

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), Article 211037. https://doi.org/10.1098/rsos.211037

Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment*, *35*(3), 189–204. https://doi.org/10.1037/pas0001200

Colombo, D., Baños, R. M., Desdentado, L., Kleiboer, A., Pavani, J.-B., Wrzesien, M., & Bretón López, J. M. (2024). Daily stress encounters: Positive emotion upregulation and depressive symptoms. *Emotion*, *24*(6), 1403–1416. https://doi.org/10.1037/emo0001362

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease*, *175*(9), 526–536. https://doi.org/10.1097/00005053-198709000-00004

Deakin, E., Ng, F., Young, E., Thorpe, N., Newby, C., Coupland, C., Craven, M., & Slade, M. (2022). Design decisions and data completeness for experience sampling methods used in psychosis: Systematic review. *BMC Psychiatry*, *22*(1), Article 669. https://doi.org/10.1186/s12888-022-04319-x

Doucette, M. M., Kwan, H., Premji, Z., Duchesne, A., Gawryluk, J. R., & Garcia-Barrera, M. A. (2024). Integration of sex/gender and utilization of ecological momentary assessment of cognition in clinical populations: A scoping review. *The Clinical Neuropsychologist*, *38*(6), 1409–1440. https://doi.org/10.1080/13854046.2024.2333579

Eisele, G., Hiekkaranta, A. P., Kunkels, Y. K., Aan Het Rot, M., Van Ballegooijen, W., Bartels, S. L., Bastiaansen, J. A., Beymer, P. N., Bylsma, L., Carpenter, R., Ellison, W. D., Fisher, A. J., Forkmann, P. D. T., Frumkin, M., Fulford, D., Naragon-Gainey, K., Greene, T., Heininga, V. E., Jones, A., . . . Kirtley, O. (2024). *ESM-Q: A consensus-based quality assessment tool for experience sampling method items*. OSF. https://doi.org/10.31234/osf.io/sjynv

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fritz, J., Piccirillo, M., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A., Norris, L., Schuurman, N. K., Snippe, E., & Bringmann, L. (2023). *So you want to do ESM? Ten essential topics for implementing the experience sampling method (ESM)*. OSF. https://doi.org/10.31219/osf.io/fverx

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

Griffith, J. M., & Hankin, B. L. (2025). Longitudinal coupling of emotional wellbeing in parent-adolescent dyads: Evaluating the role of daily life positive affect socialization processes. *Development and Psychopathology*, *37*(1), 342–360. https://doi.org/10.1017/S0954579423001633

Hakun, J. G., Benson, L., Qiu, T., Elbich, D. B., Katz, M., Shaw, P. A., Sliwinski, M. J., & Mossavar-Rahmani, Y. (2025). Cognitive health benefits of everyday physical activity in a diverse sample of middle-aged adults. *Annals of Behavioral Medicine*, *59*(1), Article kaae059. https://doi.org/10.1093/abm/kaae059

Heggestad, E. D., Kreamer, L., Hausfeld, M. M., Patel, C., & Rogelberg, S. G. (2022). Recommendations for reporting sample and measurement information in experience sampling studies. *British Journal of Management*, *33*(2), 553–570. https://doi.org/10.1111/1467-8551.12489

Himmelstein, P. H., Woods, W. C., & Wright, A. G. C. (2019). A comparison of signal- and event-contingent ambulatory assessment of interpersonal behavior and affect in social situations. *Psychological Assessment*, *31*(7), 952–960. https://doi.org/10.1037/pas0000718

Horstmann, K. T., & Ziegler, M. (2020). Assessing personality states: What to consider when constructing personality state measures. *European Journal of Personality*, *34*(6), 1037–1059. https://doi.org/10.1002/per.2266

Janssens, J. J., Kiekens, G., Jaeken, M., & Kirtley, O. J. (2024). A systematic review of interpersonal processes and their measurement within experience sampling studies of self-injurious thoughts and behaviours. *Clinical Psychology Review*, *113*, Article 102467. https://doi.org/10.1016/j.cpr.2024.102467

Janssens, K. A. M., Bos, E. H., Rosmalen, J. G. M., Wichers, M. C., & Riese, H. (2018). A qualitative approach to guide choices for designing a diary study. *BMC Medical Research Methodology*, *18*(1), Article 140. https://doi.org/10.1186/s12874-018-0579-6

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction*, *114*(4), 609–619. https://doi.org/10.1111/add.14503

Kalokerinos, E. K., Erbas, Y., Ceulemans, E., & Kuppens, P. (2019). Differentiate to regulate: Low negative emotion differentiation is associated with ineffective use but not selection of emotion-regulation strategies. *Psychological Science*, *30*(6), 863–879. https://doi.org/10.1177/0956797619838763

Kaurin, A., King, K. M., & Wright, A. G. C. (2023). Studying personality pathology with ecological momentary assessment: Harmonizing theory and method. *Personality Disorders: Theory, Research, and Treatment*, *14*(1), 62–72. https://doi.org/10.1037/per0000596

Kirtley, O. J., Eisele, G., Kunkels, Y. K., Hiekkaranta, A., Van Heck, L., Pihlajamäki, M. R., Kunc, B., Schoefs, S., Kemme, N., Biesemans, T., & Myin-Germeys, I. (2024). *The experience sampling method (ESM) item repository*. OSF. https://doi.org/10.17605/OSF.IO/KG376

Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experience-sampling methods. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920924686

Kuchibhotla, A. K., Kolassa, J. E., & Kuffner, T. A. (2022). Post-selection inference. *Annual Review of Statistics and Its Application*, *9*, 505–527. https://doi.org/10.1146/annurev-statistics-100421-044639

Kurten, S., Vanherle, R., Beullens, K., Gebhardt, W. A., Van Den Putte, B., & Hendriks, H. (2022). Like to drink: Dynamics of liking alcohol posts and effects on alcohol use. *Computers in Human Behavior*, *129*, Article 107145. https://doi.org/10.1016/j.chb.2021.107145

Lafit, G. (2022). Non-normal, higher-level, and Var(1) models for the analysis of ESM data. In I Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 185–214). The Center for Research on Experience Sampling and Ambulatory Methods Leuven (REAL).

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

Lafit, G., Revol, J., Cloos, L., Kuppens, P., & Ceulemans, E. (2024). The effect of different construct operationalizations, study duration, and preprocessing choices on power-based sample size recommendations in intensive longitudinal research. *Assessment*, *32*(2), 206–223. https://doi.org/10.1177/10731911241286868

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), Article 33267. https://doi.org/10.1525/collabra.33267

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Mestdagh, M., & Dejonckheere, E. (2021). Ambulatory assessment in psychopathology research: Current achievements and future ambitions. *Current Opinion in Psychology*, *41*, 1–8. https://doi.org/10.1016/j.copsyc.2021.01.004

Muthén, B., & Muthén, L. (2017). Mplus. In W. J. Van Der Linden (Ed.), *Handbook of item response theory* (pp. 507–518). CRC Press LLC.

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. https://doi.org/10.1002/wps.20513

Myin-Germeys, I., & Kuppens, P. (Eds.). (2022). *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (2nd ed.). Center for Research on Experience Sampling and Ambulatory Methods Leuven.

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & Van Os, J. (2009). Experience sampling research in psychopathology: Opening the black box of daily life. *Psychological Medicine*, *39*(9), 1533–1547. https://doi.org/10.1017/S0033291708004947

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A. E. G., Lewis, S. W., & Dunn, G. (2011). Experience sampling research in individuals with mental illness: Reflections and guidance. *Acta Psychiatrica Scandinavica*, *123*(1), 12–20. https://doi.org/10.1111/j.1600-0447.2010.01596.x

Perski, O., Keller, J., Kale, D., Asare, B. Y.-A., Schneider, V., Powell, D., Naughton, F., Ten Hoor, G., Verboon, P., & Kwasnicka, D. (2022). Understanding health behaviours in context: A systematic review and meta-analysis of ecological momentary assessment studies of five key health behaviours. *Health Psychology Review*, *16*(4), 576–601. https://doi.org/10.1080/17437199.2022.2112258

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., & Britten, N. (2006). *Narrative synthesis in systematic reviews: A product from the ESRC methods programme*. Lancaster University.

Postma, M. R., Vrancken, S., Daemen, M., Der Meulen, I. H., Volbragt, N., Delespaul, P., De Haan, L., Van Der Pluijm, M., Breedvelt, J. J. F., Van Der Gaag, M., Lindauer, R., Van Den Berg, D., Bockting, C., Van Amelsvoort, T., Schwannauer, M., Doi, L., & Reininghaus, U. (2024). Working mechanisms of the use and acceptability of ecological momentary interventions: A realist evaluation of a guided self-help ecological momentary intervention targeting self-esteem. *BMC Public Health*, *24*(1), Article 1633. https://doi.org/10.1186/s12889-024-19143-z

R Core Team. (2024). *R: A language and environment for statistical computing*. https://www.R-project.org/

Revol, J., Ariens, S., Lafit, G., Adolf, J., & Ceulemans, E. (2023). *Episode-contingent experience-sampling designs for accurate estimates of autoregressive dynamics*. OSF. https://doi.org/10.31234/osf.io/wbu3x

Revol, J., Carlier, C., Lafit, G., Verhees, M., Sels, L., & Ceulemans, E. (2023). *Preprocessing ESM data: A step-by-step framework, tutorial website, R package, and reporting templates*. OSF. https://doi.org/10.31234/osf.io/hnu2t

Schiavone, S. R., Quinn, K. A., & Vazire, S. (2023). *A consensus-based tool for evaluating threats to the validity of empirical research*. PsyArXiv. https://doi.org/10.31234/osf.io/fc8v3

Schneider, S., Junghaenel, D. U., Smyth, J. M., Fred Wen, C. K., & Stone, A. A. (2024). Just-in-time adaptive ecological momentary assessment (JITA-EMA). *Behavior Research Methods*, *56*(2), 765–783. https://doi.org/10.3758/s13428-023-02083-8

Scholten, S., Lischetzke, T., & Glombiewski, J. A. (2022). Integrating theory-based and data-driven methods to case conceptualization: A functional analysis approach with ecological momentary assessment. *Psychotherapy Research*, *32*(1), 52–64. https://doi.org/10.1080/10503307.2021.1916639

Schreuder, M. J., Ariens, S., Lafit, G., & Ceulemans, E. (2024). Bouncing back from emotional ups and downs: Insights in emotional recovery using survival analyses of burst ESM data. *Cognition and Emotion*, *39*(4), 910–919. https://doi.org/10.1080/02699931.2024.2423706

Seizer, L., Schiepek, G., Cornelissen, G., & Löchner, J. (2024). A primer on sampling rates of ambulatory assessments. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000656

Shi, G., Wu, Y., Pang, H., & Liu, Z. (2024). The double-edged sword effect of leaders' fear of missing out on leaders' creativity: An experience sampling method study. *Psychology Research and Behavior Management*, *17*, 259–271. https://doi.org/10.2147/PRBM.S449490

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Stieger, S., Aichinger, I., & Swami, V. (2022). The impact of nature exposure on body image and happiness: An experience sampling study. *International Journal of Environmental Health Research*, *32*(4), 870–884. https://doi.org/10.1080/09603123.2020.1803805

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, *24*(3), 236–243. https://doi.org/10.1207/S15324796ABM2403_09

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, *129*(1), 56–63. https://doi.org/10.1037/abn0000473

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, *21*(12), Article e14475. https://doi.org/10.2196/14475

Van Dalen, M., Snijders, A., Dietvorst, E., Bracké, K., Nijhof, S. L., Keijsers, L., Hillegers, M. H. J., & Legerstee, J. S. (2023). Applications of the experience sampling method (ESM) in paediatric healthcare: A systematic review. *Pediatric Research*, *95*, 887–900. https://doi.org/10.1038/s41390-023-02918-2

Van Roekel, E., Keijsers, L., & Chung, J. M. (2019). A review of current ambulatory assessment studies in adolescent samples and practical recommendations. *Journal of Research on Adolescence*, *29*(3), 560–577. https://doi.org/10.1111/jora.12471

Vankelecom, L., Schacht, O., Laroy, N., Loeys, T., & Moerkerke, B. (2025). A systematic review on the evolution of power analysis practices in psychological research. *Psychologica Belgica*, *65*(1), 17–37. https://doi.org/10.5334/pb.1318

Viechtbauer, W. (2021a). Statistical methods for ESM data. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology* (pp. 153–183). Center for Research on Experience Sampling and Ambulatory Methods Leuven (REAL).

Viechtbauer, W. (2021b). Structuring, checking, and preparing the data. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology* (pp. 137–152). Center for Research on Experience Sampling and Ambulatory Methods Leuven (REAL).

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562. https://doi.org/10.1111/j.1365-2729.2010.00368.x

Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., Vaessen, T., Kuppens, P., & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods*, *54*(6), 2981–2992. https://doi.org/10.3758/s13428-021-01777-1

Weiss, D. M., Bernier, E., Robbins, D. R., Elacqua, K. M., Johnson, K. A., Powers, K., Mesholam-Gately, R. I., & Woodberry, K. A. (2022). Using experience sampling methodology data to characterize the substance use of youth with or at-risk of psychosis. *Frontiers in Psychiatry*, *13*, Article 874246. https://doi.org/10.3389/fpsyt.2022.874246

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137

Wright, A. G. C., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment*, *31*(12), 1467–1480. https://doi.org/10.1037/pas0000685

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538