Behavior Research Methods (2025) 57:124
https://doi.org/10.3758/s13428-025-02626-1

ORIGINAL MANUSCRIPT

ESM-Q: A consensus-based quality assessment tool for experience sampling method items

Gudrun Eisele[1] · Anu Hiekkaranta[1] · Yoram K. Kunkels[2] · Marije aan het Rot[3,8] · Wouter van Ballegooijen[4,5] · Sara Laureen Bartels[6,7] · Jojanneke A. Bastiaansen[8] · Patrick N. Beymer[9] · Lauren M. Bylsma[10] · Ryan W. Carpenter[11] · William D. Ellison[12] · Aaron J. Fisher[13] · Thomas Forkmann[14] · Madelyn R. Frumkin[15,16] · Daniel Fulford[17,18] · Kristin Naragon-Gainey[19] · Talya Greene[20] · Vera E. Heininga[21] · Andrew Jones[22] · Elise K. Kalokerinos[23] · Peter Kuppens[24] · Kathryn L Modecki[25] · Fabiola Müller[26,27,28] · Andreas B. Neubauer[29] · Vanessa Panaite[30,31] · Maude Schneider[32] · Jessie Sun[33] · Stephen J. Wilson[34] · Caroline Zygar-Hoffmann[35,36] · Inez Myin-Germeys[1,37,38] · Olivia J. Kirtley[1,37,38]

Accepted: 15 February 2025 / Published online: 21 March 2025

Abstract

The experience sampling method (ESM) is increasingly used by researchers from various disciplines to answer novel questions about individuals' daily lives. Measurement best practices have long been overlooked in ESM research, and recent reviews show that item quality is often not reported in ESM studies. The absence of information about item quality may be partly explained by the lack of consensus on how ESM item quality should be evaluated. As part of the ESM Item Repository project (esmitemrepository.com)—an international open science initiative that collects ESM items in an open item bank and evaluates their quality—we brought together 42 international ESM experts to develop an ESM item quality assessment tool. In four Delphi phases, experts suggested 57 item quality criteria, rated the criteria, provided arguments for and against the criteria, and rated the criteria again, considering reflections from other experts. The result of the Delphi process is ESM-Q: a quality assessment tool consisting of 10 core criteria, as well as an additional 15 supplementary criteria, to be used depending on the type of items being rated and the availability of supplementary information. The criteria cover topics ranging from construct validity to the optimal wording of items. ESM-Q can aid ESM researchers in selecting existing ESM items, developing new high-quality ESM items, and evaluating the quality of ESM items in systematic reviews. Expert reflections also highlight open research questions surrounding ESM item design that form a research agenda for ESM measurement.

Keywords Ambulatory assessment · Ecological momentary assessment · Delphi study · Item quality criteria · Questionnaire development

Introduction

Over the last several decades, researchers from a range of disciplines have become increasingly interested in collecting self-report data in daily life with the experience sampling methodology (ESM; Larson & Csikszentmihalyi, 1983; Myin-Germeys & Kuppens, 2021; also referred to as ecological momentary assessment, Stone & Shiffman, 1994, and ambulatory assessment, Ebner-Priemer & Trull, 2009, although the use of this terminology is not consistent). ESM typically entails participants completing multiple short questionnaires per day over several days and allows researchers to gather detailed data about individuals' daily lives that can be used to answer a wide variety of research questions. ESM is currently used in psychology and adjacent fields to investigate a broad spectrum of topics such as health behaviors (Perski et al., 2022), clinical symptoms (May et al., 2018), social behaviors and experiences (Langener et al., 2023), personality dynamics (Freund et al., 2024), couple dynamics (Zygar et al., 2018), developmental processes (van Roekel et al., 2019), parenting strategies (Bülow et al., 2022), and emotion regulation/coping (Modecki et al., 2022). Compared to traditional self-report methods, ESM prioritizes ecological validity, reduces recall biases, and is particularly well suited for investigating dynamic processes as they unfold

Gudrun Eisele and Anu Hiekkaranta shared first authors.

Extended author information available on the last page of the article in individuals' daily lives (Mestdagh & Dejonckheere, 2021). However, the very nature of ESM—that it captures dynamic processes in daily life—brings new and significant methodological challenges. Measurement in ESM studies is particularly challenging, because practices common in retrospective self-report questionnaire research are often unsuitable or challenging to implement.

Measurement practices in ESM research: The status quo

As the use of ESM continues to expand, there is a high demand for validated ESM items. ESM questionnaires are generally highly specific to the goal of a particular study, but often include items to assess participants' momentary feelings, cognitions, behaviors, and context. As in traditional questionnaires, response scales for ESM items can take various forms, including Likert scales or visual analogue scales for continuous variables, single or multiple-choice response options for categorical variables, and open-text responses (Eisele et al., 2021). In addition, because ESM questionnaires need to be filled in during busy real-life situations, the use of single-item scales for constructs is common (Dejonckheere et al., 2022). Traditional self-report questionnaires that have often undergone thorough validation[1] are usually not suitable for use in ESM research due to their length and their different target constructs (e.g., traditional questionnaires tend to focus on retrospective assessment of traits, while ESM questionnaires aim to capture momentary experiences).

As a result, ESM researchers often develop new items for their specific research questions, either ad hoc or by adapting items from traditional self-report questionnaires. A review of 633 ESM studies on health behaviors observed that only a third of the studies reported including items that had been used in an ESM study before (Perski et al., 2022). Consequently, the field of ESM research has been flooded with a large variety of different items aimed at measuring the same construct (Brose et al., 2020; Heininga & Kuppens, 2021; May et al., 2018; Singh & Björling, 2019). In addition, reviews have noted that often no quantitative evidence supporting the validity or reliability of items is provided in published work. For instance, in 63 reviewed ESM articles published in three major psychopathology journals, it was found that only 30% provided information on the psychometric properties of items (Trull & Ebner-Priemer, 2020). Similarly, Hall and colleagues (2021) reviewed 234 articles investigating mood and anxiety symptomatology with ESM, covering more than 4600 items, and stated that over half of the included articles did not provide any information on the validity or reliability of items. Additionally, exact item wordings and qualitative item validity information beyond psychometric properties, such as information on response processes and possible ways of misunderstanding items that may be gathered during pilot tests, are often not shared publicly, making it difficult for researchers to build on each other's work to improve ESM items. To illustrate, a review of ESM use in physical activity research found that only about half of the 30 included articles stated the exact items, and none of them provided any information on the content validity of their ESM measures (Degroote et al., 2020).

This missing information on item quality has potentially serious consequences for ESM research, as low-quality items can produce misleading results and thereby waste scarce public resources. The lack of reporting of ESM item development and evaluation practices makes it difficult to judge the validity of conclusions from ESM studies, facilitates the propagation of low-quality items into new studies, and therefore represents an obstacle to cumulative science (for overviews of suboptimal measurement practices and lacking ESM item quality information, see for example Stone et al., 2023; Vogelsmeier et al., 2023).

Evaluation of ESM items

Theoretical and applied research on item development and evaluation to increase item validity has a long tradition in survey research. In this context, validity has been defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association [AERA] et al., 2014; but different definitions exist) and is therefore dependent on the intended use of a measure. Scholars have distinguished different types of validity evidence such as evidence based on content, response processes, the structure of a measure, and the relationships of a measure with other variables and measures (the last type is also referred to as convergent, discriminant, predictive, and criterion validity; current distinction of types of evidence following AERA et al., 2014). The reliability of a measure refers to the extent to which the measure is error-free, or produces the same outcome across repeated administration over time or with a different instrument, and affects the validity of a test. Following the above definition of validity, hereafter, we refer to the *quality of an item* as the ability of an item to measure a defined construct, with a high-quality item being an item that represents a valid measure of the construct.[2]

[1] We acknowledge that this is not always the case. See Flake and Fried (2020) for a discussion of questionable measurement practices in self-report questionnaires.

[2] We thereby assume that measuring a defined construct is the intended use of an item (as opposed to, for instance, creating a measure for purely predictive purposes). Note that we use the term item quality rather than validity, because we consider validity to be a broader term that can include aspects not addressed in the current study.

Many different methods of item development and evaluation have been described in the wider survey literature. While some of these methods have been applied to ESM items, there is currently no consensus on best practices for item development and evaluation in ESM research. Existing efforts to improve item development and evaluation in ESM research have mostly focused on quantitative validity evidence. Due to temporal dependencies, nesting of data in individuals, and the expectation that constructs change over time within individuals, the application of traditional quantitative methods to evaluate validity and reliability is not straightforward in ESM data. Consequently, researchers have developed new ways of calculating reliability and evaluating the structure of multi-item measures in ESM and other longitudinal studies (a full discussion of these methods is beyond the scope of this paper, but see Bolger & Laurenceau, 2013; Cranford et al., 2006; Muthén, 1994; Nezlek, 2017; Schönbrodt et al., 2021; Schuurman & Hamaker, 2019; Schuurman et al., 2015; Vogelsmeier et al., 2020). The routine use of single-item measures in ESM makes the evaluation of items using traditional quantitative tools especially challenging, and new strategies to overcome these challenges have been discussed (Dejonckheere et al, 2022; Schuurman & Hamaker, 2019; Schuurman et al., 2015). Consistent with this focus on quantitative assessments of item quality in the methodological literature, validity evidence in published ESM research often includes information on correlations of items with other retrospective or ESM measures (e.g., aan het Rot et al., 2015; Cloos et al., 2023; Daniëls et al., 2020) or the internal structure of measures (e.g., Chung et al., 2022; Stevens et al., 2020). However, the evaluation of items based purely on quantitative evidence is incomplete. Previous research has shown that questionnaires can have excellent psychometric properties in the absence of any meaning (Maul, 2017). In fact, many problems with items, such as basic comprehension issues, may not be detected when relying purely on such psychometric indices (Wolf et al., 2023). Indeed, researchers have repeatedly argued that different methods need to be combined to evaluate the functioning of items in a comprehensive way (e.g., Simms, 2008).

The current study focuses on expert review as a promising yet underused method of ESM item evaluation (for an exception, see Lenferink et al., 2022). Expert review refers to the evaluation of a measure by experts using a predefined set of quality criteria (Gehlbach & Brinkworth, 2011). Being relatively cost-effective compared to other validation methods, expert review is especially well suited for use early in the item evaluation process (Gehlbach & Brinkworth, 2011). Up to now, no specific tool for expert review exists for ESM research. Quality criteria for expert review have been investigated for decades for traditional questionnaires, and this research has resulted in an abundance of guidelines for item development (e.g., Haynes et al., 1995; Krosnick & Presser, 2010; Fowler & Cosenza, 2009) and even in official pretesting tools for items (e.g., the Survey Quality Predictor (Saris & Gallhofer, 2007), the Question Understanding AID (Graesser et al., 2006), and the Question Appraisal System (Willis & Lessler, 1999)). Some of these general guidelines from traditional questionnaires may be applied to ESM items, such as considerations on item wording (e.g., use simple, familiar words; use simple syntax; avoid words with ambiguous meanings, Krosnick & Presser, 2010).

However, ESM items differ from traditional items in important ways and may therefore require additional methodological considerations (Degroote et al., 2020; Horstmann & Ziegler, 2020; Palmier-Claus et al., 2011; Wright & Zimmermann, 2019). A general difference is that ESM items are administered repeatedly in the context of real life, which may amplify problems that also apply to items in traditional questionnaires. For example, it is crucial to consider the user experience of ESM items carefully, as suboptimal items may lead to participant burden and reduced data quality and quantity in intensive ESM studies (e.g., Horstmann & Ziegler, 2020; Kimhy et al., 2012). Items that may be appropriate in controlled lab environments may not be suitable to answer in noisy real-life situations. In addition, researchers must carefully consider whether a given item can function well across the whole range of situations that an individual faces during their daily lives. Further, it is important to reflect on the time frame of the item, as ESM offers many different options (e.g., “right now” or “since the last assessment”; see also Singh & Björling, 2019). Finally, the risk of triggering behavioral change or other forms of measurement reactivity by presenting certain items becomes more salient due to the repeated administration of questionnaires in ESM studies. Such differences between traditional and ESM items have been highlighted in the literature. For instance, in one of the few works on the development of ESM items, Horstmann and Ziegler (2020) have suggested that the appropriateness of the time frame and the breadth of the construct of ESM items need to be evaluated. Similarly, Palmier-Claus and colleagues (2011) provided several specific guidelines for ESM items, such as the need to assess a momentary state and to be relevant across real-life situations. However, to date, no comprehensive evaluation tool has been developed for ESM items, making it difficult to evaluate items in a systematic way.

The ESM Item Repository project and the current study

To foster transparency and method reproducibility in ESM research, the ESM Item Repository was launched in 2018. This international, open science initiative is gathering items into a publicly accessible online repository (https://www.esmitemrepository.com; Kirtley et al., 2019), which to date contains over 1000 items. The ESM item repository project comprises three work streams that aim to (I) collect ESM items in a publicly available item bank, (II) evaluate their quality, and (III) conduct large-scale psychometric validation of the items. To carry out the quality evaluation of the collected items planned in stream II of the project, the current Delphi study was set up to develop an item quality assessment tool for ESM items. The current study aimed to use the experience of ESM researchers around the world to compile a comprehensive list of ESM item quality criteria that can aid the development and evaluation of ESM items in the future. Our goal was to create a tool specifically for ESM items, because scholars have argued that they require other considerations than traditional items (Degroote et al., 2020; Horstmann & Ziegler, 2020; Palmier-Claus et al., 2011; Wright & Zimmermann, 2019). In addition, we aimed to create a user-friendly tool that could be easily applied by ESM researchers. To achieve this goal, we conducted a preregistered Delphi study during which ESM experts were invited to propose, and subsequently evaluate, quality criteria for ESM items. The Delphi technique is a tool to achieve "consensual guidance on best practices" in a field (Jünger et al., 2017) and is particularly suitable when research on which to base best practices is limited (Nasa et al., 2021). Delphi studies are heterogeneous, but typically, experts on a particular subject are invited to reflect on a selection of issues or questions across multiple rounds, with the goal of reaching consensus (Linstone & Turoff, 1975). Commonly, experts are able to see their own and others' anonymized responses from previous rounds, to facilitate reflection on the positions taken. Compared to other, related techniques such as focus groups, the relative anonymity of a Delphi process avoids distortion of the results by group dynamics (Jünger et al., 2017).

Method

Recruitment

To develop a list of potential participants, we constructed a recruitment survey in which individuals could nominate themselves or other potential ESM experts for participation in the study (see supplementary materials for the full survey; https://osf.io/pd78k). The survey assessed the inclusion criteria and inquired about the geographical location, research institution, and gender of the suggested expert to obtain a balanced sample of experts in the study. To be eligible for participation in the study, experts were required to have full professional proficiency in English. Further, experts needed to fulfill two criteria based on experience with ESM and knowledge of ESM. To fulfill the experience criteria, participants were required to hold a PhD in psychology, statistics, methodology, or a related health or social science field and to have been involved in ESM research for more than 7 years (including PhD). Based on feedback we received during recruitment, the minimum experience was lowered to four or more years and the requirement of having completed a PhD was dropped to facilitate recruitment (this represents a deviation from the preregistration). To fulfill the knowledge criterion, participants were required to have published at least one first-author paper using ESM in a peer-reviewed journal within the last three years. Experts who were direct supervisors of the researchers conducting the study were excluded from participation. A maximum of one expert per research group could participate. In cases where multiple researchers from the same research group were nominated, one of them was chosen randomly (this led to the exclusion of 10 potential participants). The recruitment survey was shared via the Twitter (now X) account of the ESM Item Repository. Additionally, the survey was sent via email to members of the Society for Ambulatory Assessment (SAA), the Belgian-Dutch Experience Sampling Method Network board, members of the Clinical Open Science mailing lists, and other key experts in the field identified by the research team but who to our knowledge were not active on social media or mailing lists. Fifty suggested experts were subsequently invited to participate in the study by email.

Procedure

Phase I

After providing informed consent, potential participants were asked to fill in a questionnaire about their experience with ESM (see Codebook in the supplementary materials for this and all other questionnaires; https://osf.io/pd78k). Experts then proceeded to the Phase I Delphi questionnaire, in which they were asked to provide a list of criteria that could be used to evaluate the quality of ESM items. Specifically, experts were asked to list what criteria they would use to evaluate the quality of an ESM item, in a situation where they knew the item (e.g., I feel sad) and the construct of interest (e.g., sadness) but did not have access to other information, such as psychometric properties. Phase I was pilot-tested in a group of ESM researchers from the Center for Contextual Psychiatry to ensure that it was understood as intended. Based on the result of the pilot, an additional explanation was added to specify that experts should list criteria relevant in a situation where psychometric properties are not known. We also asked experts to suggest a rating scale for using the criteria in practice.

The responses from Phase I were analyzed with thematic analysis, adapted from Braun and Clarke (2006, 2022). This involved the following steps:

Familiarization with the data Three members of the research team (GE, AH, OJK) read through all of the answers provided by the experts.

Generating initial codes AH and GE independently generated initial code/s for each suggested criterion; for example, the codes "clear" and "unambiguous" were created for the expert-suggested criterion "Clear and without ambiguity" by one of the coders. Codes were used to identify elements in the responses and refer to "the most basic segment, or element, of the raw data or information that can be assessed in a meaningful way regarding the phenomenon" (Boyatzis, 1998, p. 63). Codes were generated in a data-driven manner, meaning that they were based on the experts' answers.

Searching for themes AH and GE independently sorted their own codes into overarching criteria-themes and when necessary, subthemes, based on the covered aspect of item quality. For instance, the codes "clear," "easy to understand," and "comprehensible" were grouped together as the subtheme "clear/easy to understand/comprehensible" under the overarching criteria-theme "clarity" in this step by one of the coders.

Reviewing themes and subthemes AH and GE independently reviewed whether responses within the same criteria-theme/subtheme formed a coherent group. When candidate themes or subthemes were not fitting, team members independently reconsidered the criteria-theme/subtheme itself, and whether individual responses did not fit into it. Criteria themes/subthemes were then reworked by creating a new theme or subtheme or by reassigning some responses to different themes or subthemes. Next, the researchers considered all themes and subthemes and evaluated whether they accurately reflected the dataset as a whole.

Defining and naming criteria In the context of this Delphi study, themes and subthemes were intended to reflect all unique issues addressed by the criteria suggested by the experts. Therefore, in this phase, each of the two analysts either selected one criterion per criteria-theme or subtheme that best captured the issue or, if necessary, reformulated a criterion per theme or subtheme into a clear full sentence. For instance, the criterion "Is the item easy to understand for members of the target population?" was formulated for the criterion-theme "understandable" by one of the coders. If any criteria did not fit into any theme and it was not clear what issue the suggested criterion addressed, these criteria were included in Phase II verbatim, to avoid interpretative misunderstandings. The criteria that the researchers considered unclear were presented in Phase II as well, but flagged with: "This criterion was marked as unclear by the research team." In addition, criteria that were not applicable in situations where only the item and the target construct were known, but required additional information, were marked as off-topic. An exception was made for criteria that relied on information that was considered to be likely to be available in situations when item quality would be rated (e.g., criteria that related to the whole ESM questionnaire or to the study population; this represents a deviation from the preregistration).

Resolving disagreements After performing the above steps independently, GE's and AH's final lists of criteria were compared. Disagreements were documented and resolved by discussion with OJK. The whole list of criteria was discussed with the rest of the ESM Item Repository team as well, and any disagreements were resolved. Once all team members agreed with the final list of criteria provided in Phase I, we proceeded with the Delphi process. The end product of the thematic analysis was a list of 49 criteria for evaluating the quality of ESM items (Table 1).

Phase II

In the second phase, experts received a summary of the findings from Phase I in the form of a list of criteria representing all unique issues identified by experts in Phase I. Experts were then asked to judge the quality of the criteria. Per criterion, experts were asked: Is this a good criterion? (1 Disagree - 2 Mostly disagree - 3 Mostly agree - 4 Agree). If experts indicated that they agreed or mostly agreed, they were provided with the option to provide an example of an item that complied with the criterion and an example of an item that did not comply with the criterion. This optional question was adapted per criterion. For example, if an expert agreed with the criterion that an ESM item should not include jargon, they were provided with the option of giving an example of an item that did include jargon and an item that did not include jargon. Whenever the provision of an example item was considered unnecessary (e.g., for criterion 21 "The item length is between 50–80 characters") or when providing an example would require participants to write out a whole ESM questionnaire (for criteria 46–49), the questions on examples were omitted. Additionally, regardless of whether experts agreed or disagreed with a criterion, they were given the option of providing their reasoning for their judgment per criterion. In addition, experts were asked whether they wanted to include any of the criteria rated by the authors as unclear or off-topic for evaluation in the next round.

Experts were also asked to provide rating scale suggestions for the quality criteria (i.e., a scale on which the quality of ESM items could be rated using the developed criteria). Several rating scales were suggested, both for similar criteria and across all criteria. Due to the large number of suggestions in Phase I, experts could indicate whether they preferred the same rating scale across all items, and which of the suggested scales they preferred. Experts could also indicate that they preferred different rating scales for

Table 1 Quality criteria suggested by experts per rating round and percentage of agreement

No.	Core criteria	Rating 1	Rating 2	Freq.
1.	The item captures an important aspect of the construct of interest	100%	100%	21
2.	The time-frame of the item is clearly defined	91%	96%	10
3.	The time-frame in the item corresponds to the expected variability in the construct	94%	82%	4
4.	The item is measuring a construct that is expected to vary over time	91%	100%	20
5.	The wording of the item is unambiguous	89%	96%	23
6.	The item does NOT include any double negations	91%	96%	3
7.	The participant can answer the item quickly	89%	93%	4
8.	The item is short	97%	89%	11
9.	The item is worded as one simple sentence/statement/question	91%	86%	1
10.	The meaning of the item is unlikely to change over the course of an ESM study	83%	81%	1
No.	Supplementary criteria	Rating 1	Rating 2	Freq.
11.	Relevant when response scale/options include a verb: The verb tense in the response scale/options is appropriate for the time-frame of the item	100%	89%	1
12.	Relevant when study population is known: Given the study population, the item does NOT include jargon	77%	79%	10
13.	Relevant when study population is known: Given the study population, the response scale/options capture an adequate range of responses	94%	82%	2
14.	Relevant when study population is known: Given the study population, the response scale/options are appropriate for the item	94%	86%	7
15.	Relevant when study population is known: The item is easy to understand for members of the target population	100%	100%	9
16.	Relevant when study population is known: The item is appropriate for the target population	100%	93%	8
17.	Relevant when sampling scheme is known: The construct is expected to vary at the temporal level that is being sampled. (e.g. minutes, hours, days, weeks)	91%	82%	14
18.	Relevant for items that include response scales: The response scale includes anchors	91%	86%	1
19.	Relevant for items that include response scales: The response scale includes anchors that match the question	97%	96%	2
20.	Relevant for items that include response scales: The response scale is clear	97%	93%	1
21.	Relevant for composite measure: The items together cover all aspects of the construct of interest	89%	75%	2
22.	Relevant when the entire ESM questionnaire in which the item appears is known: The item is worded in the same manner as other similar questions in the ESM questionnaire	74%	70%	1
23.	Relevant when the entire ESM questionnaire in which the item appears is known: The item is sufficiently distinct from other items in the ESM questionnaire	79%	71%	2
24.	Relevant when the entire ESM questionnaire in which the item appears is known: Item ordering is unlikely to influence how participants interpret the item	74%	71%	1
25.	Relevant when the entire ESM questionnaire in which the item appears is known: Other items in the ESM questionnaire are unlikely to influence how participants interpret the item	77%	75%	1
No.	Non-agreed criteria	Rating 1	Rating 2	Freq
26.	The item assesses one construct as opposed to several constructs	71%	61%	14
27.	The item assesses only one aspect of a construct	71%	68%	1
28.	The item follows a theoretical framework	49%	35%	5
29.	The item is phrased as “in the moment” (as opposed to being retrospective)	34%	14%	6
30.	If the item is not phrased as “in the moment”, it refers to another recent time	63%	54%	2
31.	Answering the item does NOT require excessive amounts of reflection (as opposed to requiring observation / perception only)	26%	43%	5
32.	The item is phrased positively/neutrally as opposed to negatively	23%	7%	2
33.	The item length is between 50–80 characters	40%	11%	1
34.	The balance between the length of the item and the amount of detail that is provided is appropriate	72%	43%	2
35.	The item assesses a state as opposed to a response	20%	4%	1
36.	The item is NOT phrased using extreme wording	43%	21%	2
37.	The item will be interpreted identically across different samples (e.g., individuals from different backgrounds, cultures, and with different mental health histories, etc.)	77%	61%	1
38.	The item can be answered in all plausible contexts	69%	46%	2
39.	The item does NOT induce changes in thoughts or behaviors	34%	14%	1
40.	The item does NOT trigger harmful thoughts or behaviors	69%	61%	1

Table 1 (continued)

41.	The item does NOT induce socially desirable responses	77%	57%	1
42.	The item does NOT induce a learning effect	66%	42%	1
43.	The item is ethically responsible	69%	57%	1
44.	The midway point of the response scale is anchored	37%	14%	1
45.	The response scale includes at least two anchors	86%	64%	1
46.	The item results in continuous data	40%	4%	1
47.	The answer options are exhaustive	76%	57%	1
48.	The answer options are mutually exclusive	60%	36%	1
49.	The item is concrete (as opposed to abstract)	69%	39%	2
50.	The item was adequately translated from and/or to English. (e.g., through back-to-back translation)	81%	48%	1
51.	The item measures the targeted construct reliably and accurately	56%	35%	1
52.	The item is a precise measure of the construct of interest	44%	26%	2
53.	If the item refers to a broader category of states, then specific examples are given	30%	9%	1
54.	The item makes sense temporally (e.g., since the last beep)	44%	26%	1
55.	It is possible to answer to this item multiple times per day (if applicable, might not apply for evening/morning questionnaire)	41%	26%	1
56.[a]	Make a difference between the instruction formulation (can be longer) and the cue-formulation (one word)	4%	0%	1
57.	The item assesses a discrete behavior (such as "*Did you smoke a cigarette*") or a subjective state (e.g. "*I feel sad*")	22%	4%	1

Note. Sample sizes: phase I = 42, phase II = 35, phase III = 28, phase IV = 23. Criteria 1–49 were suggested in phase I and rated in phase II (Rating 1) and III (Rating 2). Criteria 50–57 were added by experts in phase II and rated in phases III (Rating 1) and IV (Rating 2). ESM/EMA/AA, referring to ecological momentary assessment and ambulatory assessment in addition to ESM in the original criteria, was replaced by ESM in the final criteria list to improve readability. Freq. = Number of experts who mentioned this criterion in their phase I suggestions. [a] This criterion was presented in phase III with a note: "This criterion was marked as unclear by the Delphi coders and is therefore presented in its original form"

different items. In this case, they were presented with all of the suggested rating scales per criterion and asked to indicate their preferred scale per criterion. The full Phase II questionnaire can be consulted in the codebook in the supplementary materials (https://osf.io/pd78k).

Phase III

In the third phase (III), we presented the anonymized list of criteria from Phase II to the same experts, along with their examples of items considered consistent or inconsistent with the criteria (see codebook in supplementary materials for the full questionnaire; https://osf.io/pd78k). The wording of criteria 1 and 31 was adapted after suggestions from experts in Phase II.[3] In addition, all criteria were reworded from questions to statements to avoid the issues with reverse coding observed in Phase II responses (12 criteria had been reverse-coded in Phase II). We also presented the experts with the reasoning provided by the other experts in Phase II. Experts were then again asked to indicate their preferred rating scale. Since most experts (71 %) were in favor of using the same rating scale across all criteria in Phase II, the five most popular rating scales were presented in Phase III, and experts could evaluate which of the scales would be most suitable to use for all criteria. At this stage, experts had the opportunity to re-evaluate their original agreement or disagreement per criterion, in light of examples and arguments provided by fellow experts. When ≥70% of experts indicated that they "agreed" or "mostly agreed" on a criterion, it was included in the final list of ESM item quality criteria. Please note that an arbitrary threshold of 70% was chosen to reflect broad agreement, and this threshold was preregistered.

Based on the suggestions from Phase II, eight additional criteria were added to the list. Experts had the chance to evaluate these criteria, explain their reasoning, and give good and bad example items (when applicable). To ensure that all criteria on the final list would have received the same number of evaluations, we added a fourth phase to the study, in which the newly added criteria could be re-evaluated.

Phase IV

In the fourth and final phase of the study, experts again had the chance to evaluate the criteria 50–57 that were added in Phase III. As in Phase III, when ≥70% of experts indicated that they "agreed" or "mostly agreed" on a criterion, it was

[3] Original formulations were: "1. Given the definition of the construct of interest, does the item capture an important aspect of it?" and "31. Does answering the item require reflection? (as opposed to requiring observation/perception only) (REVERSE-CODED: Experts stated that good items should not require reflection)".

included in the final list of ESM item quality criteria. Phase IV included the same elements as Phase III, but in relation to the suggested criteria 50–57.

Results

Participants

Of the 50 invited participants, 42 completed Phase I (54% female), 35 Phase II (61% female), 28 Phase III (54% female), and 23 Phase IV (45% female) of the study (see Fig. 1 in the supplementary materials for a flowchart of nominated and participating experts; https://osf.io/6wazb). In Phase I, experts were currently conducting research in Europe (*N* = 20), North America (*N* = 16), Australia (*N* = 5), and both Europe and Asia (*N* = 1). The average years of experience with ESM was 8 years (SD 5.44),[4] and participants had been an author on an average of 16 scientific articles using ESM (SD 24.98).[5] Participants' subjective rating of their own knowledge of ESM in terms of being on top of ESM-related literature and new developments of the method was 5.71 on a scale of 1 = a little knowledgeable to 7 = very knowledgeable. More than half of the participants had published one or more papers/book chapters/books that were primarily focused on the application of the ESM (e.g., papers on best practices or guidelines). Participants were invited to the next phase only if they had participated in the previous phase. Twenty-three participants took part in all four phases.

Criteria

In Phase I, 49 criteria covering distinct issues with the conceptual quality of ESM items were suggested (Table 1). Each criterion was rated twice, either in Phases II and III or Phases III and IV. Over two rating rounds, 25 criteria reached a consensus with over 70% of the experts indicating that they agreed with each criterion. These 25 criteria were divided by the research team into core criteria and supplementary criteria.

Core criteria

This section provides the details of our thematic analysis of the 10 criteria suggested by experts in Phase I, which resulted in the core criteria list, with excerpts from Phase I. The 10 core criteria represent criteria that are always applicable, as item wording and information about the construct the item aims to measure are generally available. We expected the elements of items assessed with these core criteria to be available in all or most cases where the quality of items is likely to be evaluated, such as when selecting items from the ESM Item Repository, when evaluating items featured in published papers, or during item development and piloting. Furthermore, in this section, we provide an overview of the experts' reflections per criterion. In these reflections, we aim to represent the diversity of thoughts and common themes discussed by the experts. Examples of good and poor items suggested by experts are provided in Table 1 in the supplementary materials (https://osf.io/z75pk). When multiple experts suggested example items for one criterion, the examples were selected based on clarity of formulation and item completeness (e.g., "*down*" was considered incomplete, while "*Sadness: I feel down*" was considered complete).

Criterion 1. The item captures an important aspect of the construct of interest Criterion 1 was formulated based on 21 experts suggesting a criterion to assess construct validity. Expert suggestions largely highlighted the importance of face validity: "*What is the face validity?*" (participant 11) and construct validity. As suggested by participant 75,[6] "*Face validity, i.e. does the item capture the construct of interest,*" we interpreted face validity and construct validity as referring to the same issue and formulated criterion 1 from participant 50's suggestion: "*Given the definition of the construct of interest, the item content reflects an apparent part of it.*"

Expert reflections In Phase II, the majority of expert reflections expressed similar views on criterion 1, e.g., "*it is almost the definition of validity*" (participant 46), "*it's the definition itself of validity*" (participant 56), with one expert commenting that criterion 1 is the most important quality criterion: "*This strikes me as the most important element. Can you recover the construct of interest from the item*" (participant 22). Experts also elaborated on why criterion 1 is important, with reference to theory: "*- if an item does not map the definition of a construct, it is not useful for theory building*" (participant 60), "*The match between item and theoretical construct is THE key quality criterion for the validity of an item*" (participant 47), while also noting that "*the challenge is that we (some scholars) sometimes oversell*

[4] Some experts reported their years of experience as years+, for example, 30+. In these cases, for the purpose of descriptive statistics, their experience was recoded as years stated, e.g., 30+ was recoded as 30.

[5] As with years of experience, when experts reported the number of authored scientific publications as number+, the number of publications was recoded as the number they stated (e.g., 15+ was recoded as 15 publications).

[6] Please note that the participant numbers were assigned based on the expert recruitment survey (before Phase I), which is why some numbers are higher than 50.

what is being tapped" (participant 15). However, despite broad agreement, a minority of experts also pointed out that judging what constitutes an "important" aspect can be subjective: "*I mostly agree, although it is somewhat subject[ive] what an 'important aspect' of a construct is, especially when this may vary from person to person*" (participant 39).

Criterion 2. The time frame of the item is clearly defined A criterion regarding a clearly defined time frame was suggested by 10 experts, and the criterion was formulated on the basis of participant 47's suggestion: "*Time frame specified in the item is clear (and defined properly) e.g., right now / since the last assessment / in the last two hours.*" Further, exactness "*The time-frame to which the item refers to is defined exactly*" (participant 50) and explicitness "*explicitness about the time frame being asked (e.g. right now, last hour, or since the last question?)*" (participant 16) were also featured in suggestions. While a time frame referring to the last assessment was brought up as an example of a clear time frame, one participant also noted that "*Since the last questionnaire could be misinterpreted"* (participant 13).

Expert reflections The majority of expert reflections illustrated that criterion 2 was considered important and necessary for all items: "*Extremely important! An ESM item should always have a time indication*" (participant 16). Experts highlighted reasons for the importance of criterion 2: "*Providing a clear timeframe is important to provide participants guidance in how to answer the question*" (participant 37), "*Time is vitally important. It is the most overlooked variable in our field. Time is embedded in every theory and every construct. If you do not account for temporality, you will never approach anything related to causality"* (participant 22). One expert also commented on the necessity of time frames in items worded in the present tense: "*However, in many cases the question will measure a current state and will be worded in present tense (I feel sad), in which case there is no time-frame. Arguably, in that case, the time frame is defined*" (participant 77). Furthermore, in contrast to the majority, six experts also suggested different approaches to defining the time frame: "*The timeframe should be defined at another level, not at the item level"* (participant 64), "*This does not necessarily need to be contained in the wording of the item though, but could also be done by e.g. introductory text*" (participant 71). Notably, wording referring to the last assessment (e.g., "*Since the last beep, I talked to someone else*") (participant 46), was featured in both "good" and "poor" example items suggested by experts.

Criterion 3. The time frame in the item corresponds to the expected variability in the construct Suggested by four experts, this criterion was formulated on the basis of participant 50's phrasing: "*The time-frame and/or wording corresponds to the expected variability of the construct.*" Participant 19 also provided an example of a match between time frame and the construct of interest: "*Does the time frame of the item wording match the time frame of the theoretical construct being assessed? (e.g. I feel sad, vs. I tend to feel sad)*" while participant 50 noted that when assessing the "*fit between wording and/or time-frame and the construct's expected variability*," an option to state "*expected variability unclear"* could be provided.

Expert reflections Criterion 3 was largely agreed to be necessary, as expressed by participant 51: "*To find reliable patterns of variability in the construct, the item has to cover the time-frame in which the change is expected to occur.*" Five experts advised against sampling too frequently "*The time frame in the item should not exceed the frequency with which the construct is expected to vary*" (participant 43), while one expert also noted that "*for some constructs it may be ideal to sample more frequently but this would cause undue burden on participants*" (participant 53). Two experts also expressed caution in the use of this criterion, given that variability of constructs can be unknown: "*I think this is important to think about. However, I'm not sure we know the answer in many cases*" (participant 39), "*I do not know how a researcher actually can use this criterion. In my view - as of yet - we know little about the (expected) variability of constructs*" (participant 75).

Criterion 4. The item is measuring a construct that is expected to vary over time The importance of assessing time-varying constructs was suggested by 20 experts, who underscored the importance of formulating items that assess fluctuating constructs: "*Make sure you use momentary and not trait-like assessment (e.g., 'I feel angry right now' and not 'I am an angry person')*" (participant 60), "*The item content refers to a (variable) state, not a (stable) trait. (this is especially important, when items are derived from classical trait questionnaires to assess corresponding states)"* (participant 50). While two participants noted that ESM would be unsuitable for stable constructs, "*Must assess a variable concept (otherwise generally unnecessary for EMA)*" (participant 40), seven participants also expressed that the expectation or potential for variance over time is important: "*Item must refer to a construct that potentially fluctuates across time*" (participant 69). Therefore, the wording for criterion 4 was derived from participant 52's criterion suggestion: "*Could reasonably be expected to vary over time.*"

Expert reflections One expert asserted that "*This is an essential criterion*" (participant 71), and the majority of arguments aligned with this notion. Further, as noted by participant 35, compliance with criterion 4 may be subject to change as new information is gained: "*If the construct is* *known to vary this is a good criterion; however, this might not yet be known for all constructs (rating might therefore need to be adapted based on new findings).*" One expert also noted that "*there may be instances where we want to demonstrate stability*" (participant 37), and in such cases criterion 4 would indeed be inappropriate to use. Another expert pointed out similar concerns: "*It could still be that I want to include an item that is not expected to vary over time. It may be that I find variability (also if I do not expect it), or I want to use the item as some kind of reference value (e.g., to identify unreliable response patterns)*" (participant 51).

Criterion 5. The wording of the item is unambiguous The issue of avoiding ambiguity was suggested by more experts than any other criterion (23 experts). The formulation provided by participant 50 was used as criterion 5: "*The wording of the item is unambiguous.*" Experts also elaborated on why clarity is necessary: "*clearly written to avoid confusion/multiple interpretation*" (participant 25). Both references to ambiguity and clarity were coded as referring to the issue of whether an item is open to multiple interpretations, given that any lack of clarity is likely a result of the possibility of multiple interpretations or the use of jargon (captured by criterion 12). Furthermore, as noted by participant 22, judging "*item clarity*" could be too subjective: "*I don't think that these are goals that can be met with rating scales.*" As evident in two experts' suggestions, the need for training or additional instructions can be an indication that an item may be ambiguous: "*How unambiguous is the interpretation of the item across participants? (vs. items that might vary in interpretation across participants, or require training for participants to understand)*" (participant 19), "*Is the item clearly formulated (e.g., is it unambiguous, were additional instructions needed for clarification)?*" (participant 71).

Expert reflections Three experts noted that avoiding ambiguity entirely is challenging, given the imprecise nature of construct definitions: "*Not always possible given some vague concepts in areas such as clinical psychology*" (participant 76), "*I think it is important to admit that ambiguity is sometimes not preventable (even if this worsens the quality of the item), especially when studying psychological phenomena*" (participant 50). Further, one expert reflected on ambiguity between and within persons. Specifically, the expert argued that participants may be allowed to have their own interpretation of an item but that this interpretation needs to be consistent within individuals: "*Some items can be interpreted on a personal level, but are coherent in one individual. However, the item still needs to be unambiguous to that one person*" (participant 65). One expert also argued in favor of ambiguity: "*Sometimes it needs to be a bit ambiguous, because we don't know the onset or end of the event or we don't know how a participant will interpret and that is okay*" (participant 15). Finally, three experts suggested the use of examples and instructions as ways to reduce ambiguity: "*this can be avoided if examples or guidance are used to improve the accessibility of items*" (participant 76).

Criterion 6. The item does NOT include any double negations A criterion to assess the presence of double negatives or complex negatives was pointed out by three experts: "*double negatives should be avoided*" (participant 18), "*Item holds no double negation*" (participant 70), and "*The phrasing does not contain complex negations*" (participant 50). Experts did not elaborate on the issue in Phase I.

Expert reflections Three experts' comments on avoiding double negatives highlighted the importance of avoiding confusion: "*This can be confusing to participants*" (participant 53). One expert elaborated on their reasoning: "*Double negations are hard for participants to understand and will hence undermine reliability of their responses*" (participant 71), and another further explained why double negatives can cause issues for participants: "*It will lead to more mistakes if there are double-[negations], as participants might not 'process' the information correctly*" (participant 65).

Criterion 7. The participant can answer the item quickly Speed as a criterion was suggested by three experts, and the criterion was formed on the basis of participant 77's wording: "*It is preferably quick to answer for a research participant.*" The same expert also referred to why items may be difficult to respond to quickly, "*Emotional items take less time than items that require thinking*" (participant 77), while another suggested an appropriate length of time a participant should be able to complete an entire questionnaire in: "*One assessment should not take no more than 1–5 minutes to complete*" (participant 76).

Expert reflections Two reasons were offered for why speed is important, namely compliance "*You will have low compliance if it takes too long*" (participant 15) and burden "*Important for reducing participant burden*" (participant 38). Nonetheless, experts also reflected on the balance between meeting research goals and burden: "*I agree, with the caveat that items that are longer would be OK if there are fewer of them (it's all about balance to reduce burden)*" (participant 60). Further, five experts commented on items that require reflection: "*Typically[,] this is important but some designs may require items that necessitate more reflection and participants may get faster over time*" (participant 53), with one expert suggesting that criterion 7 not be applied to reflective items: "*That is preferable, but this criterion is not applicable for items that might require some reflection*" (participant 47).

Criterion 8. The item is short A criterion to assess the length of an item was brought up by 11 experts. Most experts provided a brief suggestion without elaboration, e.g., "*brief/*

concise” (participant 23). Two experts also referred to reasons for the importance of shortness: “*brevity: short and concise, to limit participant burden and attention required to read and comprehend”* (participant 36), “*short so that participants will respond to it regularly*” (participant 15).

Expert reflections Several reasons for the importance of brevity were provided, related to user experience “*Short items are legible on small, portable electronic devices and thus contribute to ecological validity*” (participant 43) and burden “*helps keep participant burden low*” (participant 60). The majority of experts' comments also considered the importance of length within reason: *“Items should not be short per se. Items should be no longer than necessary to convey the content*” (participant 71). In addition, one participant offered advice for the application for criterion 8: “*I think the evaluation of [criterion 7] and [criterion 8] make only sense when comparing two items that aim to assess the same state, but not across states*” (participant 50).

Criterion 9. The item is worded as one simple sentence/statement/question One expert suggested that “*The item should be (one) simple sentence/statement/question*” (participant 75), and criterion 9 was worded on the basis of this suggestion. The research team considered this criterion distinct from criterion 8 (The item is short) because it contained an additional element regarding item structuring beyond item brevity.

Expert reflections The majority of expert reflections on criterion 9 centered on discussing situations where other formulations may be necessary: “*It's good to keep items short and clear to reduce burden. However, there are occasionally times where you might want to include an extra clarification sentence or something, and that can be valid in some instances*” (participant 60), such as when a time frame has changed: “- *sometimes you might need an extra statement, e.g. to indicate a switch in time frame”* (participant 71), or when the aim is to capture abstract constructs: “- *if you are dealing with abstract constructs then you may need one or two sentences (and an example) to ensure full understanding by participants*” (participant 40).

Criterion 10. The meaning of the item is unlikely to change over the course of an ESM study One expert provided the suggestion for a criterion that was used as a basis for criterion 10: “*Meaning is unlikely to change substantially with repeated ratings (i.e., measurement reactivity)*” (participant 52).

Expert reflections The definition of change and difficulty in assessing change in meaning was brought up in the majority of reflections: “*In general, I agree. But what is considered ‘change’? What if a person is completing EMA prompts over the course of a treatment study and their perceptions of themselves begin to shift?*” (participant 37). Experts also reflected on why assessing potential change in meaning can prove challenging: “*There is little data about measurement invariance during an ESM study, and it is plausible that many/most items change in meaning to some degree. This could be useful to weed out really extreme examples that are highly likely to change, but I think for many items it would be difficult to assess*” (participant 52).

Supplementary criteria

The 15 supplemental criteria which reached a consensus of ≥70% agreement are also shown in Table 1. These criteria were compiled into the separate supplementary criteria list by the research team, either because they only apply to specific types of items or because applying them to evaluate items requires supplementary information which may not be available to those outside the research team. Therefore, each supplementary criterion includes a qualifying statement detailing when the use of the criterion is relevant.

Non-agreed criteria

An additional 31 criteria were suggested and did not reach consensus (Table 1). For these criteria, in all cases, experts were less likely to agree with a criterion in their second round of rating compared to the first round, after assessing the reasoning and examples provided by other experts, indicating that for these 31 criteria, additional arguments moved responses toward disagreement. Expert reflections provided for criteria that nearly reached consensus (≥60% agreement; criteria 26, 27, 36, 37, 40, 45), as well as for criteria that did not reach consensus in the current study but have previously been suggested in the ESM literature (criteria 31 and 38), are highlighted in the supplementary material (https://osf.io/3sgkc) .

Off-topic criteria

In Phase I, 23 additional recommendations for ESM methodology were made (e.g., regarding the use of past research in item development, smartphone application features, psychometrics, and study characteristics). These recommendations could not be included in the criteria list because they required supplementary information unlikely to be available in most situations (e.g., information about how the item was developed, data, or details on the purpose of the study) and were therefore coded as off-topic in the context of the goals of the present study. These recommendations are available in the coding of round 1 responses in the supplementary materials (https://osf.io/z4ktq).

Response scale

In Phase II, 25 of the experts indicated that they preferred a single rating scale for all criteria, while 10 experts preferred different rating scales per criteria. In Phase III, rating scale preferences for using the quality criteria were distributed as follows: visual analogue scale ($N=3$) 1–3 Likert ($N=1$), 1– 4 Likert ($N=1$), 1–5 Likert ($N=5$), 1–6 Likert ($N=0$), 1–7 Likert ($N=7$), Yes/No ($N=0$), Yes/No/Unclear or unknown ($N=6$), and other; Likert scale with unclear or unknown option ($N=0$). Given the largely subjective nature of the final list of criteria that reached the consensus threshold of 70% agreement, we considered the Yes/No/Unclear or unknown to be the only practically useful scale. Based on reviewer comments, we have further added a field to note reflections for each criterion.

Discussion

The current Delphi study adds to the existing ESM measurement literature in two important ways. First, we developed ESM-Q—a set of quality criteria that ESM researchers can add to their toolbox and use during the development and evaluation of ESM items. Second, thanks to our consensus-based approach, the suggested criteria can be interpreted as the state of the art of measurement practices in ESM research. The criteria thereby point to important considerations and highlight gaps in the evidence base of current ESM measurement practices that need to be addressed in the future.

The scope of ESM-Q

The suggested criteria and accompanying reflections point towards topics that ESM experts currently consider priorities during item development. Many core criteria were ESM-specific, covering topics ranging from the time frames of items (criteria 2–4) to response shifts in repeated assessments (criterion 10). The applicability of some criteria was limited to certain types of items, namely items with (specific types of) response scales (criteria 11, 18, 19, and 20) or composite measures (criterion 21). However, we also observed that many suggested criteria required information beyond the item and target construct, indicating that experts consider this additional information necessary to evaluate item quality. Most salient was experts' agreement that items should be tailored to the study population, a thought that was reflected in five supplementary criteria and that also features prominently in both the ESM and the wider measurement literature (e.g., Horstmann & Ziegler, 2020). Additionally, three supplementary criteria required the availability of the entire ESM questionnaire (criteria 22–25), and criterion 17 highlighted the match between item and sampling schedule. While the target population is typically described in published research, this is not always the case for sampling schedules and ESM items, let alone entire ESM questionnaires and exact response scales. The fact that experts expect a relationship between full questionnaire characteristics, study design, and the validity of items also highlights the importance of reporting all study procedures. Therefore, we encourage researchers to follow available reporting guidelines for ESM research (e.g., Trull & Ebner-Priemer, 2020, Stone & Shiffman, 2002, van Roekel et al., 2019) and to transparently register their study plans, including a full list of ESM items, prior to data collection or access (Kirtley et al., 2021). Following such guidelines consistently would make it easier to evaluate ESM items and, consequently, the validity of the research that uses the items as well.

How to use ESM-Q

ESM-Q can be used during the development of new ESM items as well as to evaluate existing ESM items. The need to evaluate existing items may arise when analyzing items from existing datasets or when researchers wish to choose existing items to use in new data collections. Reviewers may equally wish to use ESM-Q to judge the quality of ESM items of new publications, or in the context of a systematic review. In most cases, rating the criteria will require in-depth knowledge of the literature on the target construct. For instance, to score criterion 1, the researcher needs to be familiar with the target construct, as is the case for criteria 3 and 4, which cover the temporal characteristics of the target construct.

A scoring template for ESM-Q can be found in the supplementary materials (https://osf.io/jqg78), and Table 2 displays an example use of the core quality criteria of ESM-Q for a selection of items from the ESM item repository (low-quality items were constructed for illustrative purposes). We recommend using ESM-Q to aid the critical evaluation of items. Therefore, we encourage researchers not only to score the criteria but also to provide their reasoning for their ratings to allow for incorporating nuance in the ratings and to facilitate critical discussions of items. Though we have taken great care to provide criteria that are as clear as possible, these criteria are nonetheless dependent on subjective evaluation by experts. The subjective nature of the criteria is also illustrated with the ratings in Table 2, which triggered extensive discussions, both within the author team and with the reviewers. Aside from the highly subjective nature of the ratings, an item that fails some criteria is not necessarily a poor-quality item, and items that pass most criteria may still not be good enough to be used in practice. We therefore caution against the use of ESM-Q ratings without further consideration. Instead, when an item fails a criterion, researchers should carefully consider possible

Table 2 Practical example of the application of the core criteria to rate ESM item quality

Item	Construct	1. The item captures an important aspect of the construct of interest	2. The time-frame of the item is clearly defined	3. The time-frame in the item corresponds to the expected variability in the construct	4. The item is measuring a construct that is expected to vary over time	5. The wording of the item is unambiguous	6. The item does NOT include any double negations	7. The participant can answer the item quickly	8. The item is short	9. The item is worded as one simple sentence/statement/question	10. The meaning of the item is unlikely to change over the course of an ESM study
To what extent did I, since the last beep, try to distract myself from my feelings?	Emotion regulation: distraction	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Reasoning: "Since the last beep" is not a clearly defined time window. Does this refer to the last beep that was filled in or the last beep that was scheduled?											
Right now, I feel sad	Negative affect: sadness	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Right now, I'm afraid to lose control	Psychotic experiences: loss of control	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Reasoning: Losing control is a very broad term and could refer to many different situations or experiences; only some of them would be classified as the intended construct "psychotic experiences: Loss of control." The wording is ambiguous.											
Today, I have invested in what I find important in my life	Psychological flexibility: values & committed action	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Reasoning: The terms "investing" and "what I find important in my life" could refer to many different things, making the item ambiguous. The item could require a great deal of reflection which can prevent the participant from responding quickly.											
What is your favorite ice-cream flavor?	Extraversion	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Reasoning: The item is not related to the target construct, no time-frame is mentioned and the answer to the item is not expected to vary over time											

reformulation and piloting in the case of new items, as well as measures that may be taken to help participants use the items, such as extra instructions or practicing the item more (briefing participants can improve item functioning in some cases, see Wen et al., 2021). In fact, given the rapidly evolving nature of ESM research and its expansion into increasingly more disciplines and fields, research can and often will have to use items that do not pass all core criteria (e.g., when conducting secondary analyses of existing ESM datasets). In these cases, it is up to the researcher to decide, for instance, how much difficulty, participant burden, and potential ambiguity they are willing to tolerate within their study, plan for how to deal with potential difficulties with items, and make decisions about which aspects of item quality to prioritize. For example, a researcher may be most concerned with the feasibility of their research among the study population (e.g., a population with known low compliance) and may therefore place more importance on criteria that are likely to relate to user experience, such as criterion 7 ("The participant can answer the item quickly"), criterion 8 ("The item is short"), and criterion 9 ("The item is worded as one simple sentence/statement/question"). In addition, there may be cases when a researcher purposefully uses an item that fails a particular criterion, as highlighted in the expert reflections. For instance, in some cases a researcher may want to include an item that is not expected to show variability over time (criterion 4), either to confirm its stability or to function as a data quality check.

However, failing some quality criteria may also be considered more serious than failing others, irrespective of the intended use. For instance, an item that does not meet criterion 1 ("The item captures an important aspect of the construct of interest") is unlikely to have much added value in a study that aims to measure a construct, irrespective of how the item scores on the remaining criteria.

Further research on the application of ESM-Q is crucial, and collaborations between ESM researchers are necessary for coordinating the evaluation of item quality. Given the subjective nature of the ratings, further research is needed to establish interrater agreement on the criteria. In addition, ESM-Q was not designed to serve as the sole validity evidence for items, but should be used alongside other resources (e.g., step-by-step manuals on item construction, such as Horstmann & Ziegler, 2020, or quantitative approaches to validity evidence, as discussed for instance in Bolger & Laurenceau, 2013; Cranford et al., 2006; Muthén, 1994; Nezlek, 2017; Schönbrodt et al., 2021; Schuurman & Hamaker, 2019; Schuurman et al., 2015; Vogelsmeier et al., 2020).

A research agenda for measurement in ESM studies

ESM-Q reveals a significant mismatch between what experts consider important or even necessary information when judging item quality and the information currently available in the ESM literature. In addition, despite experts' agreement on what determines item quality, the lack of an evidence base for most criteria is striking. The list of criteria can therefore serve as a research agenda that may help the field to move from consensus-based to evidence-based guidelines for ESM measurement. As such, this study situates itself within a larger movement of recent efforts to increase the validity of ESM measurements (e.g., the MITNB [Measurement is the New Black] consortium https://mitnb.org/; recent publications such as Cloos et al., 2023; Piccirillo et al., 2024; Stone et al., 2023; Vogelsmeier et al., 2023; Wright & Zimmermann, 2019). In the following, we would like to highlight a few main themes related to the proposed criteria and briefly discuss their current evidence base as well as possible approaches to extend it.

1. Construct definitions and time frames

At the very minimum, ESM researchers wanting to measure a construct should state exactly what it is they want to measure. We believe that in many instances, clear construct definitions, which are for instance needed to rate criterion 1 ("The item covers an important aspect of the construct of interest") and criterion 21 ("The items together cover all aspects of the construct of interest"), are not available. For example, even seemingly simple constructs such as "being alone" fail to be clearly defined in published work, leaving it open whether alone is defined based on the physical proximity of other people, the level of interaction, or the experienced closeness of others in a particular study. This ambiguity in construct definitions makes it hard to evaluate whether an item is tapping into the right construct. In some cases, more empirical work may be necessary to develop clear construct definitions, for instance, by using qualitative research to define the breadth of a construct in daily life (e.g., do emotion regulation items map onto patients' experiences; Stumpp et al., 2023). Relatedly, studies that use relatively high sampling frequencies (e.g., every 15 min, see Kockler et al., 2018) or that allow for assessing relationships between constructs across different time frames (e.g., Bülow et al., 2023) may help shed light on the temporal characteristics of constructs that can inform the time frame of ESM items (see criteria 2, 3, 4, and 17). Aside from the expected temporal characteristics of the construct, the exact wording of the time frame of an ESM item needs to be carefully considered, see criterion 2. Few studies have investigated the functioning of different time frames in ESM items (Boesen et al., 2018; Stone et al., 2020), and more systematic research on the consequences of using different time frames is needed to make informed decisions on the wording of items.

2. User experience

Experts generally agreed that in ESM studies, researchers need to pay special attention to the user experience when designing questionnaires, to avoid participant burden and possible negative effects on the collected data (e.g., attrition, noncompliance, careless responding). Yet, the effects of different ESM item characteristics on data quality and quantity remain largely unknown, and should be investigated systematically in the future. This includes the item characteristics reflected in the core criteria, such as answering speed (criterion 7), item length (criterion 8), and syntactic complexity (criterion 9). However, other item or questionnaire characteristics, such as the response scale or the repetitiveness of the questionnaires, may also influence the user experience, a possibility that should be investigated empirically.

3. Response pr*ocesses*

Despite being identified as a main source of validity evidence in the literature (AERA et al., 2024), there is a general lack of research on response processes for ESM items (i.e., the cognitive processes that participants engage in when responding), a gap that should be addressed in the future (Stone et al., 2023). In addition, the possibility that response processes to items change over time is unique to the repeated measures that characterize ESM (reflected in criterion 10 and criterion 42 on learning effects). Yet, stable interpretations of items are typically necessary for comparisons between people and within people over time. Whether response processes to items change over time is an empirical research question that could be approached from a quantitative perspective (e.g., whether there are systematic changes in data over time, e.g., Eisele et al., 2023; whether the measurement model changes over time, e.g., Vogelsmeier et al., 2020) or from a qualitative perspective (the changes in response processes reported by participants, e.g., Schreuder et al., 2020; Maciejewski, 2023). Now that initial empirical evidence on shifts in response processes has been published (e.g., Vogelsmeier et al., 2023), it is also important to consider what such shifts mean for the validity of ESM measures and how researchers can handle such response shifts in practical terms. Supplementary criteria 24 and 25 further raise the possibility that response processes may be influenced by adjacent items in an ESM questionnaire (such effects have been documented in cross-sectional survey research; e.g., similarly worded item pairs have been found to be rated more similarly the closer they are presented to each other in a questionnaire, Weijters et al., 2009). In ESM research in particular, researchers have recommended asking for more volatile experiences first in the questionnaire (Palmier-Claus et al., 2011), but we are not aware of any research that has explicitly investigated order effects for ESM items.

4. Response scales

The design of response scales for ESM items also featured prominently among the agreed-on criteria in the current Delphi study (criteria 18–20). Yet, the research needed to guide the choice of response scales for ESM items is currently not available. When choosing a response scale, the overall scope of the response scale needs to be considered to evaluate whether it can capture an adequate range of responses. The scope of response scales is especially relevant to consider for items that tap into extreme experiences and may result in skewed data. Related to this, new approaches to avoid skewedness by rating items in relation to previous assessment moments (Dejonckheere et al., 2023) and the use of analytical strategies to handle skewed data have been described in the recent literature (von Klipstein et al., 2023). In addition, the effects of the exact response scale format need to be evaluated in detail. This includes evaluating the use of anchors on ESM data (see research on this topic from the cross-sectional field, e.g., Weijters et al., 2010). Notably, experts did not agree on whether scales should include a midway anchor, indicating that more empirical research is needed to resolve this disagreement. Adding to this, visual analogue and Likert scales are currently used largely interchangeably in the ESM literature, but research suggests that different types of scales may result in different data, for instance in terms of the distributional characteristics of responses (Haslbeck et al., 2023) and relationships of responses with external criteria (Haslbeck et al., 2024). Yet, which type of scale should be used in which situations is currently unclear. In line with the preceding themes, this series of open questions about different types of response scales calls for more empirical research.

5. Using ESM items across populations

While ESM is increasingly used across a wide range of different clinical and nonclinical populations, the functioning of ESM items across these populations has not received much attention. Yet, many agreed-on criteria depend on the target population (12–16). Evaluating item functioning across populations is therefore an important step that ESM researchers should include in their validation efforts. Researchers could adopt both quantitative (e.g., measurement invariance analysis; see for instance Murray et al., 2020, who compared ESM item functioning across genders) and qualitative approaches (e.g., cognitive interviewing) to evaluate whether response processes differ across different target groups. In addition, including members of the target population early on during the item development phase can help create effective items (see also Schorrlepp et al., 2025).

Strengths and weaknesses

To our knowledge, this is the first Delphi study in the field of ESM research. The ESM item repository team purposefully chose this consensus-based approach instead of formulating the criteria in isolation, in order to gather unique insights into the standards of the field. The results of the Delphi process highlight important issues surrounding ESM measurement. Based on this process, we present not only a quality assessment tool but also a research agenda to work towards evidence-based measurement practices. Yet, we would like to discuss limitations and considerations for the use of the Delphi technique that may guide its use in the future. First, the Phase I input generation process was challenging, as input was provided as text, and its interpretation relied on the coders' understanding of the issues the experts highlighted. To keep the scope of the current study practically manageable, ideas discussed by experts in Phase I were coded as off-topic if they were not applicable to rating the quality of items, given the information that raters were likely to have. Therefore, valuable ideas may have received insufficient attention in the current study. Future research could use the knowledge gained in this study as a building block, for instance, by focusing on a selection of the off-topic criteria, to establish best practices for the evaluation of psychometric properties of ESM items. Furthermore, while relying on text-based and non-synchronous communication coordinated by a research team allowed the participation of a large group of experts in the current study, it also presented challenges for comprehension. The research team aimed to formulate experts' suggested criteria as clearly as possible while staying as close as possible to the suggestions made by experts. However, several arguments against criteria also noted that they did not understand a specific criterion or the wording, and asking the experts who suggested specific criteria to provide additional explanation was not feasible for the scope of the current study. Therefore, if experts struggled to understand a criterion, they were required to rely on other experts' arguments and example items for clarification. However, difficulties with criteria comprehension largely concerned criteria that did not reach consensus for inclusion in the core criteria list or the supplementary criteria list. We therefore believe that future research could also turn to the non-agreed criteria and all of the responses to find additional considerations that are relevant for ESM item evaluation.

Constraints on generality

We strove for a representative sample of experts, yet the Delphi experts were mostly based in Europe or North America; only one researcher based in Asia took part, and researchers from Africa and South America were not represented. While this may reflect the use of ESM in the literature, a major goal is to also extend ESM research to researchers in parts of the world that are currently not represented.

Conclusion

The current paper describes the use of the Delphi technique for the development of ESM item quality criteria. In four Delphi rounds, the participating ESM experts developed ESM-Q—an ESM item evaluation tool that can be applied to aid the development of new ESM items and the evaluation of existing items. ESM-Q covers topics ranging from construct validity and the time frame of items to item wording and considerations on the user experience. In addition, the Delphi process highlighted open research questions surrounding ESM measurement. We believe that addressing these open questions is crucial to improving measurement practices in ESM research.

Acknowledgements We thank all Delphi experts for their contribution to the study, including Stuart Ferguson, Eiko Fried, Laura König, Thomas Kwapil, Alexander J. Shackman, Eeske van Roekel & Mario Wenzel. In addition, we would like to thank two anonymous reviewers and the editor for helping us improve the manuscript.

Authors' contributions Gudrun Eisele: Conceptualization, Data curation, Formal analysis, Investigation (Core research team), Methodology, Project administration, Visualization, Writing—original draft Anu Hiekkaranta: Conceptualization, Data curation, Formal analysis, Investigation (Core research team), Methodology, Project administration, Visualization, Writing—original draft Yoram K. Kunkels: Conceptualization, Investigation (Core research team), Methodology, Writing—review and editing Inez Myin-Germeys: Conceptualization, Investigation (Core research team), Funding acquisition, Methodology, Supervision, Writing—review and editing Olivia J. Kirtley: Conceptualization, Formal analysis, Investigation (Core research team), Methodology, Project administration, Writing—review and editing, Supervision Remaining co-authors: Investigation (Delphi panelist), Writing—review and editing

Funding GE and AH were supported by Research Foundation Flanders (G0F8416N; G049023N; 1223725N), OJK by Research Foundation Flanders (FWO 1257821N) and KU Leuven (CPLUS/24/009), IMG and OJK by KU Leuven (C16/23/011), RWC by the National Institute on Alcohol Abuse and Alcoholism (NIAAA, K23 AA029729), SJW by National Institutes of Health grant R01DA055774, PK by IBOF grant iBOF/21/090 and KU Leuven C14/23/062, KN-G by a grant from the National Institute of Mental Health (R01MH118218), JAB by Stichting tot Steun VCVGZ (grant no. 239), VEH by the Dutch Research Council (NWO; 406.XS.04.051), LMB by the National Institute of Mental Health (R01MH118218, P50MH130957), KM by the Western Australian Future Health Research and Innovation Fund DF2022-23/1 and a Partnership Award #0059 from the Stan Perron Charitable Foundation, and ABN by the German Research Foundation (NE 2480/1-1).

Data availability The OSF page of the project contains the supplementary materials, including all materials and anonymized data (https://osf.io/pnw5f/?view_only=a21e83e6f6094053ba09928b13f29050).

Code availability Not applicable.

Declarations

Conflicts of interest No authors report conflicts of interest related to this manuscript.

Ethics approval All study procedures were approved by the Social and Societal Ethics Committee at KU Leuven, Belgium (approval number: G-2020-2064).

Consent to participate Participants provided informed consent before taking part in the study.

Consent for publication All authors have approved the submission of this manuscript for publication. Study participants have provided informed consent regarding the publication of their data in anonymized form.

References

Aan het Rot, M., Moskowitz, D. S., & Young, S. N. (2015). Impulsive behaviour in interpersonal encounters: Associations with quarrelsomeness and agreeableness. *British Journal of Psychology, 106*(1), 152–161. https://doi.org/10.1111/bjop.12070

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*.

Boesen, V. B., Nissen, S. B., Groenvold, M., Bjorner, J. B., Hegedüs, L., Bonnema, S. J., Rasmussen, Å. K., Feldt-Rasmussen, U., & Watt, T. (2018). Conversion of standard retrospective patient-reported outcomes to momentary versions: Cognitive interviewing reveals varying degrees of momentary compatibility. *Quality of Life Research, 27*(4), 1065–1076. https://doi.org/10.1007/s11136-017-1762-7

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

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology, 9*(1), 3. https://doi.org/10.1037/qup0000196

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

Bülow, A., van Roekel, E., Boele, S., Denissen, J. J., & Keijsers, L. (2022). Parent–adolescent interaction quality and adolescent affect—An experience sampling study on effect heterogeneity. *Child Development, 93*(3), e315–e331. https://doi.org/10.1111/cdev.13733

Bülow, A., Boele, S., Lougheed, J. P., Denissen, J. J. A., van Roekel, E., & Keijsers, L. (2023). A Matter of Timing? Effects of Parent-Adolescent Conflict on Adolescent Ill-being on Six Timescales. https://doi.org/10.31234/osf.io/k2d5s

Chung, J. M., Harari, G. M., & Denissen, J. J. A. (2022). Investigating the within-person structure and correlates of emotional experiences in everyday life using an emotion family approach. *Journal of Personality and Social Psychology, 122*(6), 1146–1189. https://doi.org/10.1037/pspp0000419

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

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin, 32*(7), 917–929. https://doi.org/10.1177/0146167206287721

Daniëls, N. E., Bartels, S. L., Verhagen, S. J., Van Knippenberg, R. J. M., De Vugt, M. E., & Delespaul, P. A. (2020). Digital assessment of working memory and processing speed in everyday life: Feasibility, validation, and lessons-learned. *Internet Interventions, 19*, 100300. https://doi.org/10.1016/j.invent.2019.100300

Degroote, L., DeSmet, A., De Bourdeaudhuij, I., Van Dyck, D., & Crombez, G. (2020). Content validity and methodological considerations in ecological momentary assessment studies on physical activity and sedentary behaviour: A systematic review. *The International Journal of Behavioral Nutrition and Physical Activity, 17*(1), 35. https://doi.org/10.1186/s12966-020-00932-9

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment, 34*(12), 1138–1154. https://doi.org/10.1037/pas0001178

Dejonckheere, E., Penne, I., Briels, L., & Mestdagh, M. (2023). For better or for worse? Visualizing previous intensity levels improves emotion (dynamic) measurement in experience sampling. *Psychological Assessment.* https://doi.org/10.1037/pas0001296

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ambulatory assessment: An innovative and promising approach for clinical psychology. *European Psychologist, 14*(2), 109–119. https://doi.org/10.1027/1016-9040.14.2.109

Eisele, G., Kasanova, Z., & Houben, M. (2021). Questionnaire design and evaluation. 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. 71–90). Center for Research on Experience Sampling and Ambulatory Methods.

Eisele, G., Vachon, H., Lafit, G., Tuyaerts, D., Houben, M., Kuppens, P., ... & Viechtbauer, W. (2023). A mixed-method investigation into measurement reactivity to the experience sampling method: The role of sampling protocol and individual characteristics. *Psychological Assessment, 35*(1), 68.

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

Fowler, F., & Cosenza, C. (2009). Design and Evaluation of Survey Questions. *The SAGE Handbook of Applied Social Research Methods* (pp. 375–412). SAGE. https://doi.org/10.4135/9781483348858.n12

Freund, V. L., Eisele, G. V., Peeters, F., & Lobbestael, J. (2024). Ripples in the water: Fluctuations of narcissistic states in daily life. *Personality Disorders: Theory, Research, and Treatment*. https://doi.org/10.1037/per0000650. Advance online publication.

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology, 15*(4), 380–387. https://doi.org/10.1037/a0025704

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID) - A Web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*(1), 3–22. https://doi.org/10.1093/poq/nfj012

Hall, M., Scherner, P. V., Kreidel, Y., & Rubel, J. A. (2021). A systematic review of momentary assessment designs for mood and anxiety symptoms. *Frontier in Psychology, 12*, 642044. https://doi.org/10.3389/fpsyg.2021.642044

Haslbeck, J., Ryan, O., & Dablander, F. (2023). Multimodality and skewness in emotion time series. *Emotion, 23*(8), 2117–2141. https://doi.org/10.1037/emo0001218

Haslbeck, J. M. B., Jover Martínez, A., Roefs, A., Fried, E. I., Lemmens, L. H., Groot, E. L., & Edelsbrunner, P. A. (2024). Comparing likert and visual analogue scales in ecological momentary assessment. https://doi.org/10.31234/osf.io/yt8xw

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Heininga, V. E., & Kuppens, P. (2021). Psychopathology and positive emotions in daily life. *Current Opinion in Behavioral Sciences, 39*, 10–18. https://doi.org/10.1016/j.cobeha.2020.11.005

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

Jünger, S., Payne, S. A., Brine, J., Radbruch, L., & Brearley, S. G. (2017). Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. *Palliative Medicine, 31*(8), 684–706. https://doi.org/10.1177/0269216317690685

Kimhy, D., Myin-Germeys, I., Palmier-Claus, J., & Swendsen, J. (2012). Mobile assessment guide for research in schizophrenia and severe mental disorders. *Schizophrenia Bulletin, 38*(3), 386–395. https://doi.org/10.1093/schbul/sbr186

Kirtley, O., Hiekkaranta, A., Kunkels, Y., Eisele, G., Verhoeven, D., & van Nierop, M. (2019). The experience sampling method (ESM) item repository.

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

Kockler, T. D., Santangelo, P. S., & Ebner-Priemer, U. W. (2018). Investigating binge eating using ecological momentary assessment: The importance of an appropriate sampling frequency. *Nutrients, 10*(1), 10–12. https://doi.org/10.3390/nu10010105

Krosnick, J., & Presser, S. (2010). Question and Questionnaire Design. In J. Wright & P. Marsden (Eds.), *Handbook of Survey Research* (2nd ed.). Elsevier.

Langener, A. M., Stulp, G., Kas, M. J., & Bringmann, L. F. (2023). Capturing the dynamics of the social environment through experience sampling methods, passive sensing, and egocentric networks: Scoping review. *JMIR Mental Health, 10*(1), e42646. https://doi.org/10.2196/42646

Larson, R., & Csikszentmihalyi, M. (1983). The Experience Sampling Method. In H. T. Reis (Ed.), *New Directions for Methodology of Social and Behavioral Science* (pp. 41–56). Jossey-Bass.

Lenferink, L. I. M., van Eersel, J. H. W., & Franzen, M. (2022). Is it acceptable and feasible to measure prolonged grief disorder symptoms in daily life using experience sampling methodology? *Comprehensive Psychiatry, 119*, 152351. https://doi.org/10.1016/j.comppsych.2022.152351

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The Delphi Method.* Addison-Wesley.

Maciejewski, D. (2023). *How do people decide how they feel? Response processes in Experience Sampling Method studies* [Conference Presentation]. Yearly meeting of the Belgian-Dutch Network for ESM Research in Mental Health.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives, 15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

May, M., Junghaenel, D. U., Ono, M., & Stone, A. A. (2018). Ecological momentary assessment methodology in chronic pain research: A systematic review. *The Journal of Pain, 19*(7), 699–716. https://doi.org/10.1016/j.jpain.2018.01.006

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

Modecki, K. L., Duvenage, M., Uink, B., Barber, B. L., & Donovan, C. L. (2022). Adolescents' online coping: When less is more but none is worse. *Clinical Psychological Science, 10*(3), 467–481. https://doi.org/10.1177/21677026211028983

Murray, A. L., Eisner, M., Ribeaud, D., & Booth, T. (2020). Validation of a brief measure of aggression for ecological momentary assessment research: The Aggression-ES-A. *Assessment, 29*(2), 296–308. https://doi.org/10.1177/1073191120976851

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Myin-Germeys, I. & P. Kuppens (Eds.). (2021). *The open handbook of Experience Sampling Methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies.*

Nasa, P., Jain, R., & Juneja, D. (2021). Delphi methodology in healthcare research: How to decide its appropriateness. *World Journal of Methodology, 11*(4), 116–129. https://doi.org/10.5662/wjm.v11.i4.116

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality, 69*, 149–155. https://doi.org/10.1016/j.jrp.2016.06.020

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

Perski, O., Keller, J., Kale, D., Asare, B. Y., 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

Piccirillo, M., Fritz, J., Cohen, Z. D., Frumkin, M., Kirtley, O. J., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N., Snippe, E., & Bringmann, L. F. (2024). A momentary assessment of the future of experience sampling research. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/82bnf

Saris, W. E., & Gallhofer, I. (2007). Design, evaluation, and analysis of questionnaires for survey research. *Wiley Series in Survey Methodology*, 391. https://doi.org/10.1002/9780470165195

Schönbrodt, F. D., Zygar-Hoffmann, C., Nestler, S., Pusch, S., & Hagemeyer, B. (2021). Measuring motivational relationship processes in experience sampling: A reliability model for moments, days, and persons nested in couples. *Behavior Research Methods*, 1-20. https://doi.org/10.3758/s13428-021-01701-7

Schorrlepp, L., Stadel, M., Bringmann, L.F., Hesselink, & E. S., Maciejewski, D. (2025). *Utilizing Qualitative Methods to Detect Validity Issues in Clinical Experience Sampling Methodology (ESM).* Manuscript submitted for publication.

Schreuder, M. J., Groen, R. N., Wigman, J. T., Hartman, C. A., & Wichers, M. (2020). Measuring psychopathology as it unfolds in daily life: Addressing key assumptions of intensive longitudinal methods in the TRAILS TRANS-ID study. *BMC psychiatry, 20*, 1–14.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70–91. https://doi.org/10.1037/met0000188

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n = 1 psychological autoregressive modeling. *Frontiers in Psychology, 6*, 1–15. https://doi.org/10.3389/fpsyg.2015.01038

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass, 2*(1), 414–433. https://doi.org/10.1111/j.1751-9004.2007.00044.x

Singh, N. B., & Björling, E. A. (2019). A review of EMA assessment period reporting for mood variables in substance use research: Expanding existing EMA guidelines. *Addictive Behaviors, 94*(May 2018), 133–146. https://doi.org/10.1016/j.addbeh.2019.01.033

Stevens, A. K., Blanchard, B. E., Talley, A. E., Brown, J. L., Halvorson, M. A., Janssen, T., …, & Littlefield, A. K. (2020). State-level impulsivity, affect, and alcohol: A psychometric evaluation of the momentary impulsivity scale across two intensive longitudinal samples. *Journal of Research in Personality*, *85*, 103914. https://doi.org/10.1016/j.jrp.2020.103914

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine, 16*, 199–202. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

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

Stone, A. A., Wen, C. K. F., Schneider, S., & Junghaenel, D. U. (2020). Evaluating the effect of daily diary instructional phrases on respondents' recall time frames: Survey experiment. *Journal of Medical Internet Research, 22*(2), e16105. https://doi.org/10.2196/16105

Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology, 19*(1), 107–131. https://doi.org/10.1146/annurev-clinpsy-080921-083128

Stumpp, N. E., Southward, M. W., & Sauer-Zavala, S. (2023). Do you see what i see? Researcher-participant agreement on single-item measures of emotion regulation behaviors in borderline personality disorder. *Assessment, 30*(1), 102–110. https://doi.org/10.1177/10731911211044216

Trull, T., & Ebner-Priemer, U. (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/10.31234/osf.io/eakyj

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

Vogelsmeier, L. V. D. E., Vermunt, J. K., Keijsers, L., & De Roover, K. (2020). Latent Markov latent trait analysis for exploring measurement model changes in intensive longitudinal data. *Evaluation & the Health Professions, 44*(1), 61–76. https://doi.org/10.1177/0163278720976762

Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2023). Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement. https://doi.org/10.31234/osf.io/uat5r

Von Klipstein, L., Servaas, M. N., Lamers, F., Schoevers, R. A., Wardenaar, K. J., & Riese, H. (2023). Increased affective reactivity among depressed individuals can be explained by floor effects: An experience sampling study. *Journal of Affective Disorders, 334*, 370–381. https://doi.org/10.1016/j.jad.2023.04.118

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing, 26*(1), 2–12. https://doi.org/10.1016/j.ijresmar.2008.09.003

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Wen, C. K. F., Junghaenel, D. U., Newman, D. B., Schneider, S., Mendez, M., Goldstein, S. E., Velasco, S., Smyth, J. M., & Stone, A. A. (2021). The effect of training on participant adherence with a reporting time frame for momentary subjective experiences in ecological momentary assessment: Cognitive interview study. *JMIR Formative Research, 5*(5), e28007. https://doi.org/10.2196/28007

Willis, G. B., & Lessler, J. T. (1999). Question Appraisal System - QAS-99. *Instruction Manual.* Research Triangle Institute.

Wolf, M. G., Ihm, E. D., Maul, A., & Taves, A. (2023). The Response Process Evaluation Method. https://doi.org/10.31234/osf.io/rbd2x

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

Zygar, C., Hagemeyer, B., Pusch, S., & Schönbrodt, F. D. (2018). From motive dispositions to states to outcomes: An intensive experience sampling study on communal motivational dynamics in couples. *European Journal of Personality, 32*(3), 306–324. https://doi.org/10.1002/per.2145

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open practices statement The study was preregistered prior to the start of the data collection on the OSF (https://osf.io/cmjvt/?view_only=43dd2c8e49fc4c6aa2a79eb0cfadedd2), and deviations from the preregistration are described in the supplementary materials (https://osf.io/38nvw). All data exclusions and manipulations are reported.

Authors and Affiliations

Gudrun Eisele[1] · **Anu Hiekkaranta**[1] · **Yoram K. Kunkels**[2] · **Marije aan het Rot**[3,8] · **Wouter van Ballegooijen**[4,5] · **Sara Laureen Bartels**[6,7] · **Jojanneke A. Bastiaansen**[8] · **Patrick N. Beymer**[9] · **Lauren M. Bylsma**[10] · **Ryan W. Carpenter**[11] · **William D. Ellison**[12] · **Aaron J. Fisher**[13] · **Thomas Forkmann**[14] · **Madelyn R. Frumkin**[15,16] · **Daniel Fulford**[17,18] · **Kristin Naragon-Gainey**[19] · **Talya Greene**[20] · **Vera E. Heininga**[21] · **Andrew Jones**[22] · **Elise K. Kalokerinos**[23] · **Peter Kuppens**[24] · **Kathryn L Modecki**[25] · **Fabiola Müller**[26,27,28] · **Andreas B. Neubauer**[29] · **Vanessa Panaite**[30,31] · **Maude Schneider**[32] · **Jessie Sun**[33] · **Stephen J. Wilson**[34] · **Caroline Zygar-Hoffmann**[35,36] · **Inez Myin-Germeys**[1,37,38] · **Olivia J. Kirtley**[1,37,38]

✉ Gudrun Eisele
gudrunvera.eisele@kuleuven.be

[1] Department of Neurosciences, Center for Contextual Psychiatry, Campus Gasthuisberg, Herestraat 49 ON5B bus 1029, 3000 Leuven, Belgium

[2] Faculty of Medical Sciences, University of Groningen, Groningen, the Netherlands

[3] Faculty of Behavioral and Social Sciences, Department of Psychology (Clinical), University of Groningen, Groningen, the Netherlands

[4] Department of Clinical, Neuro and Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[5] Department of Psychiatry, Amsterdam University Medical Centre, Amsterdam, the Netherlands

[6] Department of Psychiatry and Neuropsychology and Alzheimer Centrum Limburg, Maastricht University, Maastricht, the Netherlands

[7] Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden

[8] Interdisciplinary Center Psychopathology and Emotion Regulation, Department of Psychiatry, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[9] Department of Psychology, University of Cincinnati, Cincinnati, Ohio, USA

[10] Departments of Psychiatry and Psychology, University of Pittsburgh, Pittsburgh, PA, USA

[11] University of Missouri-St. Louis, St. Louis, Missouri, USA

[12] Department of Psychology, Trinity University, San Antonio, TX, USA

[13] Department of Psychology, University of California, Berkeley, CA, USA

[14] Department of Clinical Psychology and Psychotherapy, Institute of Psychology, University of Duisburg-Essen, Duisburg, Germany

[15] Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

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

[17] Sargent College of Health & Rehabilitation Sciences, Boston University, Boston, MA, USA

[18] Department of Psychological & Brain Sciences, Boston University, Boston, MA, USA

[19] School of Psychological Science, University of Western Australia, Perth, Australia

[20] Department of Clinical, Educational and Health Psychology, University College London, London, UK

[21] Faculty of Behavioral and Social Sciences, Department of Pedagogy and Educational Sciences, University of Groningen, Groningen, the Netherlands

[22] Liverpool John Moores University, Psychology, Liverpool, UK

[23] Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Australia

[24] Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium

[25] School of Psychological Science, Telethon Kids Institute, University of Western Australia, Perth, Australia

[26] Department of Medical Psychology, Amsterdam UMC Location University of Amsterdam, 1105 AZ Amsterdam, The Netherlands

[27] Amsterdam Public Health, Global Health, Amsterdam, the Netherlands

[28] Amsterdam Public Health, Mental Health, Amsterdam, the Netherlands

[29] Institute of Psychology, RWTH Aachen University, Aachen, Germany

[30] Research and Development Service, James a. Haley Veterans Hospital, Tampa, Florida, USA

[31] Department of Psychology, University of South Florida, Tampa, Florida, USA

[32] Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

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

[34] Department of Psychology, The Pennsylvania State University, University Park, PA, USA

[35] Department of Psychology, LMU Munich, Munich, Germany

[36] Department of Psychology, Charlotte Fresenius Hochschule München, University of Psychology, München, Germany

[37] Leuven Child and Youth Institute, KU Leuven, Leuven, Belgium

[38] Leuven Brain Institute, KU Leuven, Leuven, Belgium