Recommendations for reporting sample and measurement information in experience sampling studies

Article

Accepted Version

RECOMMENDATIONS FOR REPORTING SAMPLE AND MEASUREMENT INFORMATION IN EXPERIENCE SAMPLING STUDIES

Eric D. Heggestad
University of North Carolina at Charlotte

Liana Kreamer
University of North Carolina at Charlotte

Mary M. Hausfeld
University of North Carolina at Charlotte

Charmi Patel
Henley Business School

Steven G. Rogelberg
University of North Carolina at Charlotte

This version 2 February 2021 Accepted for publication in the British Journal of Management.

Not edited or formatted for publication.

Acknowledgements: We wish to thank Frederick L. Oswald for his thoughtful comments on a draft of this manuscript.

Corresponding Author: Eric D. Heggestad, University of North Carolina at Charlotte, Department of Psychological Science, 9201 University City Blvd, Charlotte, NC, 28223.

Email: edhegges@uncc.edu

Abstract

Over the last couple of decades, studies using the experience sampling methodology (ESM) have been used with increasing frequency within the management-related sciences as the method allows researchers the opportunity to investigate questions involving on-going, dynamic, intra-individual processes. Given the longitudinal nature of the methodology and the resulting multi-level data structure, there are sample- and measurement-related issues that make ESM studies different from other methods commonly used in management research. Consequently, ESM studies have demands for reporting sample- and measurement-related information that differ from more commonly used methods. In the current paper, we review the conceptual foundations of sample and measurement issues in ESM studies and report the findings of a survey of the ESM studies to identify current reporting practices. We then offer clear, easy to implement recommendations for reporting sample- and measurement-related aspects of ESM studies. We hope that these recommendations will improve reporting of ESM studies and allow readers the opportunity to more fully and comprehensively evaluate the research presented.

Keywords: Experience Sampling Methodology; Sample Attrition; Scale Adaptation; Measurement; Reliability; Reporting Standards

RECOMMENDATIONS FOR REPORTING SAMPLE AND MEASUREMENT INFORMATION IN EXPERIENCE SAMPLING STUDIES

Many management scholars are interested in relational processes, including those between specific organizational members (e.g. employees) and those between individual employees and the organization itself. These relational processes are, by their very nature, dynamic. For instance, how a follower feels about their leader is likely to change based on the actions of the leader, perhaps even on a day-to-day basis. Likewise, an employee's level of organizational commitment may change based on the issuance of a new policy or the way the organization handles a crisis situation. Conducting research to better understand how these relational processes develop and unfold is challenging. One-time questionnaire administrations to organizational members are notably deficient for studying such phenomena as this approach fails to account for the complexities of the dynamic nature of relational processes. Fortunately, promise for investigating relational processes is found in the experience sampling methodology (ESM[1]; e.g., Csikszentmihalyi, Larson, & Prescott, 1977; Csikszentmihalyi & Graef, 1980). ESM provides a powerful way for researchers to examine complex, ecologically-based research questions related to short-term, within-person dynamic processes and to explore how those processes are influenced by changes in contextual factors (Ebner-Priemer, Eid, Kleindienst, Stabenow, & Trull, 2009; Fisher & To, 2012). As such, the use of the ESM allows management scholars to directly examine how changes in contextual factors affect the moods, thoughts, perceptions, and behaviors of organizational members.

[1] There are several variants of the ESM design, including daily diary studies and the event-contingent recoding method. We have chosen to use term ESM in this paper, though our comments and recommendations are relevant to the other designs as well.

The repeated assessment of participants creates design and implementation challenges that are different from research methodologies that have been more commonly employed in management research. Fortunately, a series of methodologically-orientated papers have been written to help researchers navigate decisions for designing and implementing an ESM study – e.g., see Beal (2015), Bolger, Davis & Rafaeli (2003), Fisher and To, (2012), Gabriel, Podakoff, Beal, Scott, Sonnentag, Trougakos and Butts (2018), Hektner, Schmidt, and Csikszentmihalyi (2007); Ohly, Sonnetag, Niessen and Zapf (2010), Scollon, Kim-Prieto and Diener (2009), and Uy, Foo, and Aguinis (2010). These papers provide excellent discussions and recommendations for researchers looking to design and carry-out an ESM study.

The unique elements of the ESM also require authors to provide different, or, in some cases, more elaborated information about aspects of their study than they might when writing about research based on other designs. This is particularly true around descriptions of the sample and of the measures given, as these are clear points of departure from more commonly employed methods. Yet, there has not been a movement toward establishing a set of reporting standards for ESM studies to date. We hope to begin that movement in this paper. Establishing reporting standards for ESM research would be valuable for at least four reasons, including (1) enhanced transparency of the science, (2) providing necessary information for future meta-analytic work, (3) understanding and evaluating replication studies, and (4) an overall improvement in the quality of our science.

In this paper, we provide recommendations for the reporting of sample and measurement-related information in manuscripts describing ESM research. Our reporting recommendations are informed by sampling theory, measurement theory, best practice recommendations for carrying out ESM studies (e.g., Gabriel et al., 2018; Ohly et al., 2010), and discussions of how to adapt, or

change, scales to fit particular research contexts (e.g., Heggestad, Scheaf, Banks, Hausfeld, Tonidandel, & Williams, 2019). Our recommendations were further informed by comparing these conceptual foundations for reporting sample- and measurement-related information with current reporting conventions within the ESM literature. Specifically, we examined a sample of 110 papers (118 individual studies) published in the management-related journals that employed an ESM design to identify current reporting conventions – i.e., what information was typically being reported in describing samples and measures. (A complete description of our methodology for collecting these reporting convention data as well as descriptive information for the sample can be found in an online appendix/supplement.) Considering these reporting conventions with respect to the conceptual foundations allowed us to identify areas where essential information was either presented ambiguously or not consistently reported. Our recommendations are constructed to carefully capture the information identified in our conceptual review and to correct or clarify issues we found in our survey of current reporting conventions. As such, the purpose of this paper is not to advise researchers on how to best conduct ESM research, but rather to provide practical, easily implemented recommendations for the nature and type of sample- and measurement-related information that should be reported in manuscripts describing ESM research.

Our paper is structured in the following way. First, we provide a discussion of the distinctions between reporting conventions, recommended reporting standards, and formal reporting standards. Second, we look at issues related to sample reporting. Here, we bring together the conceptual foundations regarding what sample information should be reported and with what we identified as reporting conventions in the ESM literature. Building on this information we provide a recommendation for a table that captures key sample-related

information, highlighting the elements of the table through the presentation of a detailed fictitious example. Third, we discuss two measurement-related issues, the adaptation of scales for use in ESM studies and reliability estimation for the resulting multi-level data. We again bring together the conceptual foundations and current reporting conventions related to each of these issues to offer reporting recommendations. We provide a sample appendix for clearly reporting scales that are adapted for use in the study, again using a fictitious research project.

Reporting Conventions, Recommendations, & Standards

Distinctions can be drawn between reporting conventions, recommended reporting standards, and formal reporting standards. *Reporting conventions* represent cultural norms within a field or disciple regarding how authors structure a manuscript and for the nature of information to be included in the manuscript. There are numerous reporting conventions in management-related research. Among others, there are norms for how long a manuscript should be (when journals don't provide page or word limits), how long an introduction section should be relative to the rest of the paper, what information should be included when describing the sample of participants, and that control variables can described and justified in far less detail than focal measures. Such culturally determined conventions vary considerably across disciplinary boundaries.

Recommended reporting standards represent informed and argued perspectives about what information should be included in a manuscript and/or the level of detail provided. Such recommendations, which are sometimes referred to as best practices, are typically offered by researchers within the field and are most typically focused on a particular methodological or analytical issue. Credé and Harms (2015), for example, provided recommendations for reporting the results of confirmatory factor analyses. Likewise, Herman Aguinis has authored a series of

papers in which he and his co-authors have included recommendations for reporting details on a variety of methodological and analytical techniques (see, for example Aguinis & Bradley, 2014; Aguinis, Gottfredson, 2010; Aguinis, Gottfredson, & Joo, 2013).

In contrast to conventions and recommendations, *formal reporting standards* are explicitly stated expectations or rules about what specific information should be reported in a journal article and, in some cases, how that information should be reported. Formal reporting standards are provided by an authoritative group or organization. The weight of the authority behind the issuing body leads to accountability and widespread adoption of the standards by authors, reviewers, and editors.

Perhaps most well-known among management scholars, the American Psychological Association (APA) published the Journal Article Reporting Standards (JARS; American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) to establish formal reporting standards for all of their journals in which quantitative methods are used. Updated and extended in 2015 (Applebaum et al., 2018),[2] the JARS cover all of the typical aspects of quantitative manuscripts, including specific reporting requirements for the abstract, the methods section, the results section, etc. For example, with regard to reporting on the sample the JARS indicate that authors are to report the "percentage of sample approached that actually participated" (Appleaum et al, 2018, p.6). With regard to the descriptions of the measures used the standards indicate that, among other things, authors should "Estimate and report values of reliability coefficients for the scores analyzed" (Appleaum et al, 2018, p.7).

[2] APA also formed a working group to establish reporting standards for qualitative research.

The establishment of reporting standards is important for at least four reasons (APA, 2008; Kazak, 2018). First, thorough reporting allows readers to make a careful, well-informed determination of the strengths and weaknesses of the research. As Applebaum et al. (2018) note, establishing reporting standards helps to ensure transparency in science, allowing readers to fully "understand the content of the report and evaluate the credibility of the results and conclusions" (p. 5). Fundamentally, reporting standards improve the likelihood that readers get the information they need to evaluate both the internal and external validity of the findings. The need for reporting transparency and reporting accuracy has become even more essential given the increasingly strong push for evidence-based decision making, which has "placed a new emphasis on the importance of understanding how research was conducted and what it found" (APA, 2008, p.2)

Second, reporting standards help to ensure that necessary information is provided in research papers to allow for careful and thorough meta-analytic work. The lack of consistent reporting in the literature for purposes of meta-analysis were key drivers leading to the issuances of two sets of formal reporting standards in the medical field: the Consolidated Standards of Reporting Trials (CONSORT; Begg et al., 1996) and the Transparent Reporting of Evaluations with Non-randomized Designs (TREND; Des Jarlis, Lyles, & Crepaz, 2004). Third, reporting standards improve the capability of researchers to conduct high-quality replication studies. Importantly, when replication fails, clear reporting of the initial study can allow researchers to identify points of divergence in the sample characteristics or the methodology which might explain the failure. Fourth, reporting standards can, over time, improve the overall quality of the science. That is, the adoption of standards within a discipline serves as a signal to researchers that careful reporting is essential for good research. Knowing that such information will need to

be thoroughly reported, researchers may make better, or at least more thoughtful, decisions as they design their research.

To date, no reporting standards for ESM studies have been issued. Given the multilevel nature of the data, the longitudinal element of the design, and the frequent need to modify measures for use in ESM studies, elements of reporting an ESM study differ in important ways from other research designs more commonly used. Below, we argue for and offer recommendations for reporting sample- and measurement-related aspects of ESM studies. We hope these recommendations might lead to improved, more thorough reporting of ESM research.

Sample Reporting

Background Information and Survey Findings

ESM studies typically require participants to complete multiple surveys over a number of time intervals. This design leads to several issues related to sample reporting. First, the data are multi-level, such that scores on the scales included in the questionnaires are nested within persons. The repeated, within-person observations reside at Level 1, while the individual (and any other between-person observations) reside at Level 2.[3] The design, therefore, requires reporting of the number of observations at both Level 1 and Level 2. Second, the longitudinal nature of the design requires reporting about both sample attrition – i.e., how many participants drop out of the study and who they are – and participant non-compliance – i.e., how many of the assessment opportunities did each participant complete and how many did they miss.

Level 2 Sample Size. Reporting the number of people (or organizations, or groups, etc.) that participate in a research study is a long standing and widely adopted reporting standard.

[3] Higher level grouping variables may also exist. For example, within-person responses (Level 1) may be nested within individuals (Level 2), which are, in turn, nested within a particular work team (Level 3). Other Level 3 grouping variables might include participant sex, supervisor, department, organization, etc.

Researchers clearly recognize the importance of providing sample size information, as it is crucial for interpreting the outcomes of statistical analyses, for evaluating the generalizability of findings and, among other factors, for subsequent meta-analytic research. As such, it is not surprising that we found that authors reported Level 2 sample size information in all of the 118 studies we examined in our survey of current reporting conventions.[4]

While a Level 2 sample size was consistently reported, it was not always clear exactly what the number provided represented. That is, it was not clear whether it reflected the number of individuals who agreed to participate in the study or the number of participants on which the analyses were based (i.e., after cases were removed from the data due to excessive missing data or having provided poor quality data). We examined a subset of our studies to better determine what the number represented. Specifically, we compared the Level 2 sample size reported in the *Sample* section to statements of Level 2 sample size in other parts of the paper, such as the numbers of participants reported in table notes. In cases where there was some ambiguity in the reported Level 2 sample size, there did seem to be a tendency for the number of participants reported to represent a cleaned, final sample. The bigger picture point here, however, is that readers should not have to dig so deep into a paper to make what amounts to an educated guess about what the reported Level 2 sample size represents. The fact is, both numbers should be reported. The JARS report clearly indicates that authors should report the number of people who began the study and the number of people on which analyses were conducted – see Table 2 and Figure 2 in Applebaum et al. (2018).

Attrition and Removal Due to Non-Compliance. ESM studies demand a lot from participants. Studies which assess participants multiple times a day over the course of several

[4] The mean number of Level 2 observations across the 118 studies we coded was 104.8 (SD = 5.92) with a range from 34 to 341.

days are intrusive and can be tedious for participants. Signals to respond to an assessment may arrive when the participant is not available or when they simply cannot take time to respond to the assessment (such as when driving a car). The demands placed on participants by these studies will inevitably result in some level of participant attrition and assessment non-compliance. Participant attrition occurs when a participant simply discontinues participation in the study (i.e., they stop responding to assessments), while participant non-compliance occurs when a participant fails to respond to one or more assessment opportunities or fails to respond to all of the items within a particular assessment opportunity (Ohly et al., 2010). Importantly, both attrition and non-compliance result in missing data. When enough assessment data are missing, researchers typically remove these cases from the data set. For example, a researcher may decide to drop from the data set those participants that completed fewer than 30% of the assessments. As such, this is a Level 2 sample issue.

When authors establish a standard for missing data and use that standard to remove cases with excessive missing data from the data set, they need to report the decision rule and the number of participants that are removed from the data (Gabriel et al., 2018; Scollon et al., 2009). This information is important as high rates of non-compliance and/or high attrition can raise concerns about non-response bias (e.g., Baraldi & Enders, 2010; Enders, 2010; Rogelberg & Stanton, 2007).

In our examination of reporting conventions, we found that 67 of the 118 papers (56.8%) discussed dropping cases from the data because of missing data. We are unsure if the other 45% of studies did not drop any cases or if they did but failed to report doing so. Sixty of the 67 (89.6%) papers that reported dropping cases because of missing data also provided information about the decision rule they had established. The rules that authors used to remove cases from

the data varied widely, in part due to design differences. For example, studies that used matched pairs (e.g., employees and supervisors) employed rules that were based on percentages of times that both individuals responded. When rules to remove participants from the data were based on the percentage of assessments completed, we saw considerable variability in the percentages of assessments that needed to be completed to be retained in the data. In one study (identifying information withheld), for example, participants were retained if they completed more than 8% of the assessments administered. In another study (identifying information withheld), participants had to complete more than 60% of the assessments to remain in the data set.

Unfortunately, we found reporting regarding the issue of removing cases because of missing data to be notably deficient in our sample of papers. While some authors carefully and precisely reported the number of cases removed and clearly specified the adjusted Level 2 sample size, this was the exception rather than the norm. In fact, we attempted to code how many participants were removed and for what reasons, but the reporting was so inconsistent and confusing that we concluded that we could not extract high quality data. Thus, a reporting convention has not emerged about whether to and how to report this information.

Level 1 Sample Size. The Level 1 sample size represents the number of assessments completed by participants over the course of the study protocol. Careful reporting of the exact Level 1 sample size is important for many of the same reasons that it is important to report the Level 2 sample size: e.g., for interpreting the outcomes of statistical analyses, for evaluating the generalizability of findings, and for secondary data analysis. Given that ESM studies can include multiple assessments per day over many days, it is not uncommon to see studies with large Level 1 sample sizes. Across the studies we examined, we found Level 1 samples ranging from 121 observations to over 6,000 observations, with a mean of 915.26 (SD = 950.41) observations.

These sample sizes are often large enough to raise potential concerns that the analyses are actually over-powered. As Gabriel at al. (2018) noted, “The possible problem with large Level 1 samples is that a significant *p* value can be detected even when the variance explained in the outcomes of interest is quite small” (p. 975-976). ESM study authors need to provide readers with accurate and precise Level 1 sample information so that readers can form an evaluation of the practical meaningfulness of the observed effect sizes.

We found that authors were by and large diligent in presenting the number of Level 1 observations. We did, however, find 15 studies (of 118, 12.7%) for which the number of Level 1 observations was not reported. This should not happen. In our discussion of Level 2 sample sizes above, we found some ambiguity around what the number reported represented. While we did not get the same sense of ambiguity in looking at the Level 1 information, we may have developed a false sense of security. That is, although we saw no statements indicating that authors had removed individual Level 1 observations due to factors such as poor data quality, the fact that it wasn’t reported doesn’t mean it didn’t happen. In fact, we expect that it probably should happen. That is, as we will discuss in greater detail below, authors should be screening the individual assessment responses and removing responses when there are high levels of missing item-level data or evidence for poor quality data.

Recommended Sample Reporting Standards.

Our review of sample reporting conventions indicates that authors are generally providing information about Level 1 and Level 2 sample sizes, though typically not at the level of detail needed for a clear understanding of what that numbers represent. To rectify this issue, we call for a standardized and clear method for reporting sample information. To this end, we propose that sample information for ESM studies be presented in a table. We provide a

representation of such a table as Table 1. We have filled in the values in the table with completely fictitious data for demonstration purposes. For this fictitious data, let us consider that we have conducted an experience sampling study in a particular company that has 300 employees and that all employees were invited to participate in the study. For the sake of this example, participants were asked to complete 3 surveys per day over the course of 5 days. In the remainder of this section we will walk through the information presented in the table and provide some discussion around the elements we include.

Level 2 Information. The top section of the table captures information about the Level 2 sample. The population size is reported in first line of this section of the table. Such information may not always be available or relevant. For example, if a researcher uses personal connections to get a sample of 75 working adults, the population would consist of all working adults that the researcher knows or could reach. In this situation, it would not be possible to define the number of people in this population, so the researcher could indicate "not applicable" on the population line of the table or simply not include that line. The population is, however, identifiable in many situations. For instance, if all of the employees in the accounting function of a particular company were offered the opportunity to participate in the study, the population would be readily known and the number of employees in the accounting function should be added to this first line in the table. In our fictitious situation, all 300 employees in the organization were invited to participate in the study. Thus, the population size is recorded as 300.

Next, we recommend that authors report the number of people in the initial sample, which we define as the number of people that completed the informed consent protocol or otherwise began participation in the study. In our hypothetical situation, we indicate that the initial sample includes 175 people, which is 58.3% of the population. While we see reporting the

size of the population as an optional piece of information, reporting the size of the initial sample is essential. This information provides a baseline by which readers can judge the final Level 2 sample; concerns about the data are likely to grow as the final sample becomes more discrepant from the initial sample.

Authors should also clearly report all cases that are removed from the data. Going one step further, we argue that they should indicate the specific reasons cases where removed and how many were removed for that reason (Aguinis, Hill, & Bailey, 2019). As we discussed above, two key reasons authors might remove cases from the data include attrition and excessive non-compliance. Of course, both attrition and non-compliance result in missing data (at Level 1). Our examination of the ESM literature made it quite clear that authors tend to focus on the fact that data are missing, and not the reasons why the data are missing (i.e., attrition vs. non-compliance). Authors who addressed missing data at all tended to set a single standard for the amount of allowable missing data, then removed people from the data set who did not meet that standard.

However, data missing due to attrition and data missing due to non-compliance may be fundamentally different. Some degree of non-compliance is almost a certainty in ESM studies. As suggested above, a signal indicating it is time to complete an assessment can arrive at a time when a participant is not available to complete it – such as when driving a car, in a meeting, or having an important phone call. By the time the participant is available to complete the assessment, the time window for the assessment may have closed. In these cases, it seems likely that the data are missing due to the circumstances in the participant's life at the time and not necessarily about the nature of what is being assessed (unless the study is, for example, about how busy the person is). As such, it seems likely that these data can be described as missing

completely at random (MCAR; Baraldi & Enders, 2010; Enders, 2010). If that is the case, then the missingness of the data is not likely to influence or bias statistical results.

In contrast, data missing due to attrition are likely to be the result of an active choice on the part of the participant to stop completing any additional assessments. This is a more complicated form of missing data. If the choice to no longer participate is due to reasons unrelated to the concepts being studied, then the data can likely be considered missing completely at random (MCAR) and would not be expected to meaningfully bias the statistical results. However, if the decision to stop participating is systematically associated with concepts under study, then the missing data could best be considered missing not at random (MNAR) and could lead to biased statistical results. Consider, for example, a study examining variability in job satisfaction. If a participant were to choose to leave the study because they are simply "fed up" with their organization, the missingness of these data would be directly related to the focal construct of job satisfaction. In this case, attrition would be systematically related to a key study variable (MNAR) and could lead to biased statistical results. Attrition could have more indirect effects as well. Given the demands of participating in ESM studies, it is possible that people with, for example, lower levels of conscientiousness may be less likely to complete the study (Scollen et al, 2003). Even if conscientiousness is not directly under study, it may be related to variables that are under study, such as job performance – i.e., if those lower in conscientiousness are more likely to leave the study, it could result in indirect range restriction on variable such as performance (Hunter, Schmidt & Le, 2006).

Of course, researchers are unlikely to know why participants chose to stop responding and if those reasons are related to the variables being studied. But, given the fact that missing data due to attrition may be more likely to result in biased results than are missing data due to

non-compliance, authors should present the number of cases dropped for each of these reasons (Applebaum et al., 2018).

The identification of who was removed due to attrition and who was removed due to excessive non-compliance can be made by an examination of the response patterns in the data. We recommend that authors first establish a rule for excluding cases for missing data – for example, removing participants who do not complete at least 30% of the assessments. Once those cases are identified, authors should examine the response patterns for those participants to determine which ones belong to each of the categories. Cases that are due to attrition will have a clear point at which the participant stopped responding. Cases that are due to excessive non-compliance will have at least occasional responses throughout the duration of the study. Researchers will have to make some judgement calls in making these decisions.

In our sample table, we show that 40 cases where removed from the data, representing 22.9% of the initial sample. The table also shows that of these 40 cases, 15 were removed due to attrition and 25 were removed for excessive non-compliance, which represent 8.6% and 14.3% of the initial sample, respectively. Importantly, and consistent with the JARS and CONSORT reports, we include a table note that indicates that excessive missing data was defined as a within-person response rate of less than 30% (Applebaum et al., 2018; Des Jarlais et al., 2004). Authors may have other reasons for removing cases from the data, such as poor data quality (which we address in greater detail below). Dropping cases for any other reasons should also be detailed in the table.

The final Level 2 sample represents the set of participants that will be included in the analyses, calculated as the number of participants in the initial sample less the number of participants that were removed from the data set. Together, this information provides the reader

with a detailed accounting of the sample – i.e., who is in, who is out, and why they are out. Readers can use this information to gauge concerns about the potential for bias and the generalizability of the findings. In the case of our hypothetical table, we see that 22.9% of cases were removed from the data. The fact that only about one-third of the cases removed were because of attrition and because those removed for attrition were only about 8.6% of the initial sample, the potential for bias due to missing data seems fairly unlikely (Heggestad, Rogelberg, Goh & Oswald, 2015). For further reading on the issues related to missing data in ESM studies as well as methods for dealing with missing data, readers are referred to Beal (2015) and Gabriel et al. (2018).

If authors want to go a step further and provide readers with a stronger opportunity to evaluate the sample and the potential impact of dropping cases, they could present key demographic information for the population (if available), the initial sample (this information would need to be collected early in the protocol) and the final sample. Of course, such information would be useful for identifying ways in which the final sample may differ from the initial sample or the population. Research by Silvia, Kwapil, Eddington and Brown (2013) found that women had higher individual-level response rates in ESM studies than did men. As such, men may be more likely to attrite from the study or to be removed from the final sample because of excessive non-compliance. Presenting demographic information for the initial and final samples would allow readers to clearly see and, consequently, evaluate the extent to which there may be concerns about changes in demographic representation within the study.

In our sample table, we provide hypothetical demographic information for sex and race. The table shows that the population is 48.0% female and 56.0% White, 31.0% Black and 13.0% Hispanic. Other demographic data could also be presented, depending on the nature of the

sample. For example, if sampling employees from a particular company, it could be helpful to present data on the functional areas in which they work (e.g., finance, HR, IT) or the level of the participant's role in the organization (e.g., entry level, individual contributor, manager, executive). Authors almost universally provide demographic information for their final samples (although, again, it is not always clear if the information presented is for a final sample or an initial sample), but also providing this information for the initial sample and/or the population could go a long way to evaluating the extent to which the final sample is representative of the initial sample or the population.

Level 1 Information. The middle section of our sample reporting table describes the Level 1 sample. The first line is the "Potential Level 1 sample size". This is simply the number of surveys that were administered over the course of the study. In our hypothetical case, we had final Level 2 sample of 135 participants who were asked to complete 3 surveys per day over the course of 5 days. Thus, our potential Level 1 sample size is 135 x 3 x 5 = 2,025[5]. The next line is the number of assessments that participants actually completed. In our hypothetical example, participants completed 1,448 of the assessments, which represents an overall response rate of 71.5%.

There are many reasons a researcher may need to remove some of the assessments from the initial Level 1 sample. One reason for removing individual assessments is due to missing item-level responses. Participants may get interrupted in the middle of responding to a particular assessment, for example, and not respond to all of the items. Depending on the number of items responded to, a researcher may choose to score the scales based on the items that the participant

[5] Since participant attrition isn't typically identified until after the study protocol ends, the possible Level 1 sample size can be calculated using the number of participants that begin the study protocol – perhaps those that complete the informed consent process.

did complete or, if the participates completed only a small to moderate number of the items, the researcher could choose to drop this assessment from the data. When authors make this latter choice, they should report how many assessments they removed from the data and the criteria they used for excluding assessments. Our sample table indicates that 16 Level 1 assessments were removed from the data and that six of these cases were due to missing item-level data, representing only .4% of all assessments. Also, a table note is provided indicating that assessments were dropped when the participant provided response to fewer than 75% of the items (note, we are not recommending this cut-off as a rule of thumb).

A second reason that researcher may choose to drop an assessment is because of evidence that the data provided are of poor quality. Over the last decade or so, survey researchers have grown increasingly wary and concerned about insufficient effort responding, which was defined by Huang, Bowling, Liu and Li (2015) as the condition under which "a person responds to items without sufficient regard to the content of the items and/or survey instructions" (p. 828). When participants engage in insufficient effort responding their responses do not reflect the attribute the item is designed to assess, and, consequently, the score on the scale (as a combination of item responses) will not accurately reflect the person's standing on the attribute. When enough participants engage in this behavior – or, in the case of an ESM study, if a participant engages in this behavior on enough of the assessments – it can have implications for the psychometric properties of the measures and the relationships between measured variables (e.g., DeSimone, DeSimone, Harms & Wood, 2018; Huang, Curran, Keeney, Poposki & DeShon, 2012; Johnson 2005; McGonagle, Huang, & Walsh, 2015; Merritt 2012). Identifying specific cases where a participant is likely to have engaged in insufficient effort responding can be difficult, but several

indices can help researchers identify those cases (e.g., Dunn, Heggestad, Shanock & Thielgard, 2016; Meade & Craig, 2012).

We did not see, in our examination of the ESM literature, any cases were researchers reported screening their data for insufficient effort responding. Although it is recommended that ESM researchers keep their assessments short (e.g., Fisher & To, 2012; Hektner et al., 2007; Ohly et al. 2010; Uy et al., 2010), which should help combat insufficient effort responding, it is still possible that respondents will engage in this behavior and, as such, the data should be screened for it. For example, a participant who is busy at the time of a signal but who wants to be compliant (perhaps because an incentive for a high completion rate was provided) may simply respond to all of the items with the same response (e.g., a 3 on a five-point Likert-type scale) without giving any attention to the items. In cases were a researcher suspects that a participant engaged in insufficient effort responding for a particular assessment, they should consider dropping that assessment. Of course, authors should report the number of assessments they drop due to concerns about the quality of the data. In our sample table, we show that 10 assessments, or 0.7% of the initial Level 1 sample, were removed due to data quality concerns.

There may be other issues that researcher encounter which lead them to decide to drop particular assessments. Those should also be reported in the table. Our sample table shows that after dropping 16 assessments the final Level 1 sample was 1,432 assessments, or 70.7% of the potential Level 1 sample size.

The last section of the table, which also captures Level 1 information, reports the frequencies of individual response rates. For instance, our hypothetical table shows that 10 participants completed between 80 and 89.9% of the assessments administered to them. Note that we only report individual response rates for those in the final Level 2 sample, thus 30-39.9% is

the last line in this section of the table since participants with response rates lower than 30% were deemed to have excessive missing data and were removed from the sample. The information in this portion of the table is helpful for providing readers with a greater perspective on how the missed assessments were distributed in the sample. If the majority of a study's participants completed less than half of all assessments administered, for example, readers may have reservations about the conclusions of the paper based on the individual response rates. In our hypothetical example, the table shows that rather few people had near perfect response rates (i.e., above 90%) and the most common individual response rate was in the 60-69.9% range.

We recognize that our proposed sample reporting table is quite large. We further recognize that including such a large table can pose issues, particularly when seeking to publish the research in an outlet that has strict page or word limitations (though tables don't always count against such limits). However, we argue that presenting this level of detail in the table can reduce the amount of text needed to fully describe the sample, and, therefore, may not add as much length to the paper as it may appear. Additionally, writing within such page limitations always require authors to make decisions about what information is of enough value to be included in the manuscript – in editing our own manuscripts we are often faced with deciding to endeavor to streamline the introduction, cut some details from the method or results section, or to remove the discussion of a particular issue in the discussion section. We believe that it is imperative to our science to fully and carefully describe the methodology; that methodology should be prioritized over other elements of the manuscript. Ultimately, a full and complete understanding of the methodology is essential for evaluating both the internal and external validity of the research. Thus, we feel that the size of table is reasonable given the amount of detailed information it provides and the importance of the information.

Measurement Reporting

Background Information and Survey Findings: Scale Adaptation

Heggestad et al. (2019) found scale adaptation – i.e., when authors change some aspect of an established measure for use in their study – to be an exceedingly common practice in management research. Specifically, they found that 81.4% of the more than 250 articles that they examined included one or more adapted scales, and that 45.8% of the more than 2,000 scales they examined were reported by authors as having been adapted in some way. In our examination of ESM studies, we found the rate of author disclosed scale adaptation to be even higher. Specifically, we found that 111 of the 118 (94.1%) studies in our sample included at least one adapted scale[6]. Further, of the 522 scales we coded from the 118 studies, authors reported having adapted 364 (71.2%) of them.[7] We were dependent on authors to disclose that they had adapted a scale. Given that Heggestad et al. (2019) found significant underreporting of author-identified scale adaptation, it is quite likely that the rate of scale adaptation we observed in our sample of studies is an underestimate of the true rate of scale adaptation.

There are many specific ways that authors alter scales. We identified 10 different forms of author-reported scale adaptation in our sample of studies. The frequency with which we saw each of these forms of adaptation are shown in Table 2. The last row of the table indicates that we observed descriptions of 32 scales (8.8% of all adapted scales) in which the authors indicated that they adapted the scale but did not provide *any* information about what aspect of the scale

[6] We coded 522 scales from the 118 studies. On average, studies included 4.42 scales, with a range from 1 to 9 scales. Of these scales, 448 (85.8%) were existing measures (or adapted from existing measures) and 74 (14.2%) were created specifically for the study. The average number of items per scale was 4.75, with a range of 1 to 21 items. Of the 522 scales coded, 114 (22.1%) were three-item scales and over half (57.0%) consisted of four or fewer items.

[7] Authors reported adapting the scales themselves in 316 of the 364 (86.8%) cases and using a scale that was previously adapted in 47 (12.9%) cases. For the remaining case, the authors noted the scale was adapted but did not provide further information.

that they changed. This lack of information is simply not acceptable. Reviewers and editors must be vigilant and, when they suspect that a scale has been adapted, demand that authors articulate the ways in which the scale was changed. Far and away, the two most common forms of scale adaptation were shortening the scale (61.5% of all adapted scales) and changing the timeframe (36.5% of all adapted scales). We briefly discuss each of these below.

Adapting Scale Length. ESM scholars have frequently called for the use of scales that are as short as possible (e.g., Fisher & To, 2012; Hektner et al., 2007; Scollon et al., 2009; Uy et al., 2010). For instance, Fisher and To (2012) suggest that surveys should not exceed 5–10 minutes, and Uy et al. (2010) suggest two minutes or less should be sufficient. Likewise, Ohly et al. (2010) indicate that "scales consisting of five or more items are usually not suitable" and suggest that, "Preferably, abbreviated and adapted scales as well as single items are used" (p. 85-86). The central issue here, as noted by Uy et al. (2010), is to "strike a balance between obtaining enough information and not overburdening participants" (p. 39). Keeping questionnaires brief should lead to higher response rates and lower rates of attrition.

While there are certainly advantages to using shortened scales in ESM research, there are also potential drawbacks. In their research examining the practice of scale adaptation, Heggestad et al. (2019) surveyed a group of psychometricians and editorial board members about concerns they had regarding various forms of adaptations. Shortening a scale for inclusion in an ESM study was rated as one of the most concerning forms scale adaptation. The concern with shortened scales is that they can lack reliability and validity. The Spearman-Brown prophesy makes clear that scales with more items should be more reliable. And, of course, with fewer items, it is less likely that the content of the items will cover the breadth of the construct (i.e., content validity). Authors have often used factor analytic information to choose the subset of

items to include in the shortened scale, selecting those items with the highest factor loadings. This approach will tend to privilege reliability over validity. Ultimately, the optimal set of items to include in a shortened scale will be those that maximize the shortened scales reliability, content validity, and convergent validity. The bottom line here is that while there are clear reasons to do so, authors should take particular care when shortening scales.

A recent paper by Cortina et al. (2020) presents a scale shortening tool that is available on the web. The tool, which requires pilot data, allows authors to specify the number of items that they want to include on a shortened scale (e.g., choose four items from a 10 item scale) and then, using the pilot data, calculates psychometric information for all possible combinations of scales of that length. With the appropriate pilot data, the psychometric information available can include information about coefficient alpha, the part-whole correlation, a convergent validity coefficient, a divergent validity coefficient, and content validity. This tool should prove quite valuable for ESM researchers looking for optimal ways to keep scales brief.

Adapting the Timeframe. Many scales used in management-related research have been designed to measure constructs that are conceptualized as relatively stable beliefs, attitudes and behaviors. As such, instructional sets of for these measures often ask respondents to consider their thoughts, feelings and behaviors "in general". In the context of an ESM study, however, the authors are generally more interested in the thoughts, feeling and behaviors of the participant over a relatively short amount of time – e.g., over the last few minutes, since the last signal, over the morning hours. Consequently, authors often need to change the timeframe that respondents consider when responding to the items, which is often done by providing timeframe information in the instructional set, the items, or both. Illies and Judge (2002), for example, modified the

items of two general job satisfaction scales so that the words “at this very moment” were added to the items.

We found changes to the timeframe to be the second most common form of adaptation in our sample of ESM studies. Specifically, authors reported changing the timeframe in 36.5% of all of the scales that we examined. The concern with altering a scale’s timeframe is that it can change the psychometric properties of the scale (Shrout, Lyons, Dohrenwend, Skodol, Solomon, Kass, 1988), and, ultimately, the construct being assessed (Zuckerman, 1983). Additionally, item content may not be equally construct relevant across different timeframes. For example, an item such as “I like to go to parties” may be a good indicator for the trait of extraversion, but an adaptation of the item, such as “I would like go to a party right now,” may not be a good indicator of state extraversion. The editorial board members and psychometricians in the Heggestad et al. (2019) study considered changes to the timeframe of a scale to fall between “a problem” and “a slight problem.” As such, changes to the timeframe of a scale should be done carefully and with consideration of the implications for the nature of the construct and the extent to which the adapted items are good indicators of the construct in the new timeframe.

Recommendations for Reporting Scale Adaptations

As Heggestad et al. (2019) suggest, “Transparency around scale adaptations is necessary for issues of replicability, for integration across studies (i.e., meta-analysis), and for evaluating the quality of the research” (p. 2613). Editors, reviewers, and readers need to know precisely how an adapted scale differs from the original, validated scale. Thus, authors need to be transparent about the changes they make, no matter how small these changes may be. To provide clear and transparent reporting of scale adaptations, we call for authors to provide an appendix

that includes detailed information about the original and adapted versions of the scales included in the study.

Considering our fictitious experience sampling study, let us consider that we used an adapted version of Owens, Johnson and Mitchell's (2013) expressed humility scale.[8] An example of the appendix detailing this adaptation is provided as Sample Appendix A. The appendix includes the citation to the original paper in which the scale development and initial evidence to support the reliability and the validity of the scale are presented. The appendix also includes a clear statement of how the scale was adapted; in our fictitious case, we indicate that the scale was shortened, that the timeframe was changed, and that the referent was changed. Next, we provide information about the instructional set provided to the respondents. As shown, Owens et al. (2013) did not provide details about the instructions that they developed with the scale, but we provide the (hypothetical) instruction set used in our fictitious study. Providing the instructions in our fictitious situation is particularly important because it is through the instructions that the timeframe adaptation occurs; participants are directed to consider their behavior over the last hour.

The appendix also provides detailed information about the response scale. Researchers change response scales very frequently. Heggestad et al. (2019) found that authors reported changing the response scale for only 4.8% of the more than 2,000 scales they coded. However, when they compared the response scale used in the validation studies to the response scales reported by researchers (Study 2), they found changes to the response scale occurred in more than 60% of scale administrations. Heggestad et al. speculated that researchers may not consider changes to the response scale to be meaningful enough to warrant mention in their papers. But

[8] Any similarity between our example and an actual experience sampling study in the published literature is entirely coincidental.

such changes do warrant mention. The response scale is an integral component of a scale and the scores that result from it. There is evidence that alterations to the responses scale can change how people respond to the items and, subsequently, the score they get on the scale (e.g., Schwarz, Knauper, Hippler, Noelle-Neumann & Clark, 1991). Thus, changes to the response scale should be reported. In our sample appendix, we provide the information that Owens et al. (2013) provided about the response scale used in their study and provide also the response scale used in our fictitious study.

Finally, both the original and the adapted sets of items should be presented (when the items are not proprietary). In our example, we adapted the scale by including only a subset of the original items and by changing the referent (from "this person" to "I"). These changes can be clearly seen by comparing the items from the original scale to those of our adapted version. In the context of scale shortening, it is important to include all of the items from the original scale so that readers can make an evaluation of the content of the shortened scale vis-à-vis the full scale.

Though not strictly a reporting issue, we feel that it is important to mention that authors need to provide evidence for the validity of adapted scales. Although a detailed discussion of this issue is beyond the scope of this paper, we do provide a brief discussion of the issue in the online appendix.

Estimating and Reporting Reliability

Reporting estimates of reliability for measured variables is exceedingly important. Reliability estimates provide information about the degree of error variance in a set of scores, as such, it provides information about the extent to which a person's observed score is an accurate indicator of the that person's true score (assuming the test is valid). Reliability estimates are also

important because reliability – or lack thereof – attenuates observed relationships between constructs (i.e., attenuation due to unreliability). Knowing not only the observed correlation between two variables, but also having estimates of the reliability for each of those variables allows meta-analytic researchers to estimate the relationship between the constructs by correcting for measurement error. Consequently, reporting estimates of reliability is widely regarded as a reporting standard. For example, the JARS statement indicates that authors should "Estimate and report values of reliability coefficients for the scores analyzed" (Applebaum et al., 2018, p. 7), and several journals (e.g., *Journal of Applied Psychology*; *Journal of Management*), in their own statements of reporting requirements, indicate that authors should provide estimates of reliability for all of their measures.

In our look at reporting conventions for ESM studies, authors frequently reported reliability estimates for the scales used in their studies. Specifically, some estimate of reliability was reported for 435 of the 522 (83.3%) scales we coded. Of the 87 cases for which no reliability was reported, 46 were single-item scales, two were said to be formative constructs, and one was scored as a factor score (though we would note reliability could still be reported for these scales, just not Cronbach's alpha). No reliability information was presented for the remaining 38 scales (7.3%). Although reliability estimates were reported in the vast number of cases, the fact that no reliability information was observed for 7% of the scales in our sample is problematic.[9]

For those scales for which reliability estimates were provided, the vast majority were Cronbach's alpha internal consistency reliability estimates (381 of 435; 87.6%). Cronbach's alpha was presented in different ways, however; authors reported a point estimate 318 times, a range of values 95 times, and both a point estimate and a range 44 times. Of the 318 occasions

[9] Detailed information about the reliability estimates, including the mean point estimates and ranges by scale length, is provided in Part 3 of the online appendix.

when the authors provided a point estimate of alpha, 163 cases were described as an average alpha – where alpha was calculated for each administration of the scale and then averaged across those administrations. For the remaining 155 cases, it was unclear how the estimate reported had been calculated (it could have been an average or calculated on a single administration of the scale). For the ranges, authors calculated alpha on each administration of the scale and reported the highest and lowest values observed.

While it is good to see that most authors are reporting reliability information for most of their scales, it must be noted that they are not doing so in the most appropriate manner. The nested nature of ESM data requires a multilevel consideration of reliability (Beal, 2015; Gabriel et al, 2018; Ohly, 2010). The calculations of alpha we just reviewed consider exclusively Level 1 data. However, for measures that are given on multiple occasions (i.e., Level 1) to the same individual (i.e., Level 2), calculating a reliability estimate on data from a single level (i.e., Level 1) confounds variance associated with each level. As noted by Geldhof, Preacher, and Zyphur (2014), "single-level reliability estimates therefore do not necessarily reflect true scale reliability at any single level of analysis" (p. 72). Multilevel reliability estimation procedures isolate these sources of variance and produce estimates of reliability both within-groups and between-groups. See Geldhof et al. (2014) and Shrout and Lane (2011) for more information about multi-level reliability estimation.

We found 15 instances (2.9% of the scales examined) where the authors reported using a multi-level reliability estimation procedure (these cases came from 3 studies). A good example was the paper by Bidee, Vantilborgh, Pepermans, Willems, Jegers, and Hofmans (2017), who provide a nice description of the multi-level approach to reliability and report within- and between-person reliability estimates for each of their scales. In their analysis of the team

inclusion scale they used, they initially found that while the between-person reliability estimate was good ($\omega_{between}$ = .93), the within-person estimate was rather low (ω_{within} = .45). Examining the results of the multilevel confirmatory factor model indicated that one of the items did not load well on the latent factor. Removing this item from the scale resulted in a notably higher level of within-person reliability for the scale.

The small percentage of scales for which multi-level reliability was evaluated, while dismal, represents an improvement over what Gabriel et al. (2018) found in their examination of the ESM literature – they found no examples of evaluations or multi-level reliability. ESM researchers need to move away from reporting point estimates and ranges of Cronbach's alpha and toward multilevel approaches to reliability estimation.

Summary & Conclusions

ESM is a powerful method that provides scholars with the opportunity to explore interesting questions about dynamic relational processes. As with the development of any method, it takes time for both best practices regarding the implementation of the method and standards for reporting the research to emerge. In this paper, we have focused on the reporting of ESM studies, particularly with regard to reporting information about the sample and the measures used. Our look at reporting conventions within the organizational science-related ESM literature found that while authors are providing much of the essential information about their samples, there is considerable variation in what authors chose to present and the clarity of their presentation. To help provide added clarity in reporting sample-related information we have provided a table that includes key information for both the Level 1 and Level 2 samples. This table provides clear information regarding participant flow through the study, the number of individuals removed from the data set and why they were removed (i.e., attrition or excessive

non-compliance), the number of individual assessments completed, and identification of how many and why some assessments were dropped from the data set.

With regard to measurement, we considered issues of scale adaptation and reliability estimation. Our look at the ESM literature unsurprisingly showed that authors frequently adapt scales for use in research, most often shortening scales and/or changing the timeframe which participants are to consider in making responses to the items of those scales. Unfortunately, in many cases, authors fail to fully describe the ways in which they adapt measures, which is certainly a concern for the transparency of our science. To help improve that transparency, we provide a sample appendix that authors can use to clearly and thoroughly describe the ways in which they adapt scales. Also with regard to measurement, we discuss that the multi-level nature of ESM data requires that authors move away from reporting point estimates and ranges of Cronbach's alpha and toward multilevel approaches to reliability estimation.

We hope that the discussion of these issues and the recommended tools for reporting sample information and scale adaptations might lead to improved transparency, replicability and completeness in the reporting of ESM studies. Such improvements in our science are important for the consumers of ESM research, allowing them the information they need to fully evaluate the internal and external validity of the research.

References

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*(4), 351-371.

Aguinis, H., Hill, N. S., & Bailey, J. R. (in press). Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods,* published online ahead of print. https://doi.org/10.1177/1094428119836485

Aguinis, H., & Gottfredson, R. K. (2010). Best-practice recommendations for estimating interaction effects using moderated multiple regression. *Journal of Organizational Behavior*, *31*(6), 776-786.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270-301.

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839-851.

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

Baraldi, A. N., & Enders, C. K. (2010).An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5-37.

Beal, D. J. (2015). ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior, 2*(1), 383-407.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin I., et al. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association, 276*, 637–639.

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

Bidee, J., Vantilborgh, T., Pepermans, R., Willems, J., Jegers, M., & Hofmans, J. (2017). Daily motivation of volunteers in healthcare organizations: relating team inclusion and intrinsic motivation using self-determination theory. *European Journal of Work and Organizational Psychology*, *26*(3), 325-336.

Cortina, J. M., Sheng, Z., List, S. K., Keeler, K. R., Kattell, L. A., Schmitt, N., Tonidandel, S, Sommerville, K., Heggestad, E. D., & Banks, G. (2020). From alpha to omega and beyond: A look at the past, present, and (possible) future of reliability assessment in the Journal of Applied Psychology [monograph]. *Journal of Applied Psychology*, *105*(12), 1351-1381.

Credé, M., & Harms, P. D. (2015). 25 years of higher-order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. *Journal of Organizational Behavior*, *36*(6), 845-872.

Csikszentmihalyi, M., & Graef, R. (1980). The experience of freedom in daily life. *American Journal of Community Psychology*, *8*(4), 401-414.

Csikszentmihalyi, M., Larson, R., & Prescott, S. (2014). The ecology of adolescent activity and experience. In *Applications of Flow in Human Development and Education* (pp. 241-254). Springer, Dordrecht.

Des Jarlais, D. C., Lyles, C. & Crepaz, N. (2004) Improving the reporting quality of non-randomized evaluations of behavioral and public health interventions. The TREND statement. *American Journal of Public Health, 94*, 360–366.

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology An International Review*, *67*(2), 309-338.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*(1), 105-121.

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T. J. (2009). Analytic strategies for understanding affective (in) stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology*, *118*(1), 195-202.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

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

Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2019). Experience sampling methods: A discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods*, *22*(4), 969-1006.

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

Heggestad, E. D., Rogelberg, S., Goh, A., & Oswald, F. (2015). Considering the effects of nonresponse on correlations between surveyed variables: A simulation study to provide context to evaluate survey results. *Journal of Personnel Psychology*, 14 (2), 91-103.

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management*, *45*(6), 2596-2627.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). Experience sampling method: Measuring the quality of everyday life. Thousand Oaks, CA: Sage Publications.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299-311.

Huang, J. L., Curran, P. G., Keeney, J. Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99–114.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594-612.

Ilies, R., & Judge, T. A. (2002). Understanding the dynamic relationships among personality, mood, and job satisfaction: A field experience sampling study. *Organizational Behavior and Human Decision Processes*, *89*(2), 1119-1139.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39,* 103–129.

Kazak, A. E. (2018). Journal article reporting standards. *American Psychologist*, 73, 1–2.

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2015). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology: An International Review, 65,* 287-321.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17,* 437–457.

Merritt, S. M. (2012). The two-factor solution to Allen and Meyer's (1990) affective commitment scale: Effects of negatively worded items. *Journal of Business and Psychology, 27,* 421-436.

Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research. *Journal of Personnel Psychology*, *9* (2), 79-93.

Owens, B. P., Johnson, M. D., & Mitchell, T. R. (2013). Expressed humility in organizations: Implications for performance, teams, and leadership. *Organization Science, 24*(5), 1517-1538.

Rogelberg, S. G., & Stanton, J. M. (2007). Introduction: Understanding and dealing with organizational survey nonresponse. *Organizational Research Methods, 10*(2), 195-209.

Schwarz, N., Knauper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, F. 1991. Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly,* 55: 570-582.

Scollon, C. N., Prieto, C. K., & Diener, E. (2009). Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being* (pp. 157-180). Springer, Dordrecht.

Shrout, P. E. & Lane, S. P. (2011). Psychometrics. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for Studying Daily Life* (pp. 302-320). London, England: Guilford Press.

Shrout, P. E., Lyons, M., Dohrenwend, B. P., Skodol, A. E., Solomon, M., & Kass, F. (1988). Changing time frames on symptom inventories: Effects on the Psychiatric Epidemiology Research Interview. *Journal of Consulting and Clinical Psychology, 56*(2), 267-272.

Silvia, P.J., Kwapil, T.R., Eddington, K.M., & Brown, L.H. (2013). Missed beeps and missing data: Dispositional and situational predictors of non-response in experience sampling research. *Social Science Computer Review, 31*(4), 471-481.

Uy, M. A., Foo, M. D., & Aguinis, H. (2010). Using experience sampling methodology to advance entrepreneurship theory and research. *Organizational Research Methods*, *13*(1), 31-54.

Zuckerman, M. (1983). The distinction between trait and state scales is not arbitrary: Comment on Allen and Potkay's "On the arbitrary distinction between traits and states." *Journal of Personality and Social Psychology, 44*(5), 1083–1086.

Table 1. Proposed sample reporting table with fictitious data

Level 2					% Female	% White	% Black	% Hispanic
Population[1] *(if applicable)*	300				48.0%	56.0%	31.0%	7.0%
Initial Level 2 sample[2]	175		58.3%	of population	43.7%	52.4%	33.7%	9.1%
Cases removed	40		22.9%	of initial Level 2 sample				
Attrition		15	8.6%	of initial Level 2 sample				
Excessive non-compliance[3]		25	14.3%	of initial Level 2 sample				
Add other reasons								
Final Level 2 sample	135		77.1%	of initial Level 2 sample	44.6%	53.9%	32.4%	8.6%
Level 1								
Potential Level 1 sample size	2,025							
Initial Level 1 sample	1,448		71.5%	of potential Level 1 assessments				
Assessments removed	16		1.1%	of initial Level 1 assessments				
Missing item responses		6	0.4%	of initial Level 1 assessments				
Data quality issues		10	0.7%	of initial Level 1 assessments				
Add other reasons								
Final Level 1 sample	1,432		70.7%	of potential Level 1 assessments				

Frequencies of Individual Response Rates					
100% completion	3	2.2%	60-69.9%	29	21.5%
90-99.9%	9	6.7%	50-59.9%	23	17.0%
80-89.9%	14	10.4%	40-49.9%	20	14.8%
70-79.9%	21	15.6%	30-39.9%	16	11.9%
				135	100.0%

Notes. [1] The population incudes all employees at XYZ Corporation. [2] Initial sample is defined as the number of people that completed the informed consent process. [3] Excessive non-compliance was defined as a within-person response rate of <30%.

Table 2. Frequency of various forms of scale adaptation

Forms of Adaptation	Count	Percent of Adapted Scales (n = 364)
Shorten the scale	224	61.5%
Change the timeframe	133	36.5%
Combined items from multiple scales	23	6.3%
Changed context	19	5.2%
Translated	15	4.1%
Changed the wording	6	1.6%
Added items	3	0.8%
Dropped items	2	0.5%
Changed referent	2	0.5%
Changed the scale anchors	1	0.3%
Unknown	32	8.8%

Note: Some of the scales were adapted in more than one way.

Sample Appendix A

Expressed Humility

Owens, B. P., Johnson, M. D., & Mitchell, T. R. (2013). Expressed humility in organizations: Implications for performance, teams, and leadership. *Organization Science*, *24*(5), 1517-1538.

ADAPTATIONS

Shortened the scale, changed the timeframe, changed the referent

Original Scale	***Adapted Version***
INSTRUCTIONS	
None provided	Indicate the extent to which each of the following statements describe your behavior over the last hour.
RESPONSE SCALE	
5-point, 5 = Strongly agree	1 = strongly disagree; 2 = disagree; 3: neither agree nor disagree; 4= agree; 5 = strongly agree
ITEMS	
1. This person actively seeks feedback, even if it is critical.	
2. This person admits it when they don't know how to do something.	
3. This person acknowledges when others have more knowledge and skills than him- or herself.	
4. This person takes notice of others' strengths.	1. I noticed others' strengths.
5. This person often compliments others on their strengths.	2. I complimented others on their strengths.
6. This person shows appreciation for the unique contributions of others.	
7. This person is willing to learn from others.	
8. This person is open to the ideas of others.	3. I was open to the ideas of others.
9. This person is open to the advice of others.	4. I was open to the advice of others.

ONLINE SUMMPLEMENT

PART 1: SURVEY OF REPORTING CONVENTIONS IN THE MANAGEMENT-RELATED ESM LITERATURE

To develop an understanding of the current reporting conventions in the management-related ESM literature, we conducted a systematic search of ESM articles published between 2013 and 2018 in seven management-related journals from the U.K., Europe, and the U.S.: *Academy of Management Journal*, *European Journal of Work and Organizational Psychology*, *Journal of Applied Psychology*, *Journal of Business and Psychology*, *Journal of Management* and *Journal of Occupational and Organizational Psychology*. Each of these journals had published at least three ESM studies in the five-year period of time we examined. We conducted a search through EBSCO Host (searching multiple databases including Business Source Complete, PsychINFO, and PsychARTICLES) for articles that included the terms "experience sampling", "ESM", "diary", or "diaries" in the abstract or as keywords. These search terms resulted in the identification of 121 articles (between 2013 and 2018). We reviewed each of those articles and excluded those that did not report empirical data derived from the ESM. We coded 110 articles, including 118 individual studies (some articles contained more than one ESM study). Given the way that we identified the papers to examine, our sample should be seen as a representation of recent ESM work in the organizational sciences, not as a comprehensive examination of the ESM literature in general.

To ensure we captured relevant and accurate information from each study, a standardized coding form was developed. The articles were coded by two members of our author team; each article was coded independently by one of the two coders. Once all articles were coded, a third

author reviewed the database to look for missing or unusual (e.g., a very large sample size; a large number of items were given) data. When missing or unusual data were identified, the third team member consulted the original article to ensure the information was accurate or to correct the code. Only a small number of codes were changed in this review process.

Descriptive information for the 110 articles is presented in the Table 1.1 below. The number of articles differ considerably across the journals we examined. Three distinct assessment intervals were present in our data: event contingent, daily, and weekly. For event contingent intervals, participants are asked to respond to an assessment every time a particular event occurred. In studies using a daily interval, assessments are administered at least once a day, often multiple times per day. Studies with a weekly interval administered assessments on a weekly basis. Descriptive information regarding the assessments for studies using each of these intervals is shown in Table 1.2 (in the table, “interval” refers to a day for studies using a daily interval and a week for studies using a weekly interval). The table shows that the vast majority of studies used daily intervals. The table also provides information about the number of questionnaires administered over the course of the study, the numbers of scales included in the questionnaires, the number of items per questionnaire and the number of items per scale.

Table 1.1 Descriptive information for coded articles

	Articles Coded	Percent of coded articles
Journal		
AMJ	18	16.4%
EJWOP	39	35.5%
JAP	24	21.8%
JBP	4	3.6%
JOM	7	6.4%
JOOP	18	16.4%
Lead Authors		
US	38	34.5%
Non-US	72	65.5%
Business School	56	50.9%
Non-Business School	54	49.1%
Article Information		
Included Appendix	12	10.9%
No Appendix Included	98	89.1%
Research was Funded	31	28.2%
Not Funded	79	71.8%
Dissertation	1	0.9%
Not a Dissertation	109	99.1%

Note: *n* = 110 articles. AMJ = *Academy of Management Journal*; EJOWP = *European Journal of Work and Organizational Psychology*; JAP = *Journal of Applied Psychology*; JBP = *Journal of Business and Psychology*; JOM = *Journal of Management*; JOOP = *Journal of Occupational and Organizational Psychology*.

Table 1.2. Assessment information for the three ESM assessment intervals

	Daily (n = 103)			Weekly (n = 7)			Event Contingent (n = 8)		
	Mean	Range: Low	Range: High	Mean	Range: Low	Range: High	Mean	Range: Low	Range: High
Intervals	8.39			8.57			N/A		
Questionnaires / Interval	2.15	1	6	1.29	1	3	1		
Total Questionnaires Given	18.50	3	80	9.14	5	25	Varies by participant		
Scales / Questionnaire	4.51	1	9	3.43	2	5	4.13	3	7
Items / Questionnaire	21.33	4	84	18.14	3	27	15.75	10	25
Items / scale	4.90	1.00	16.00	5.62	1.00	12.00	3.88	3.33	4.67

Note: Interval is how often the assessments were given; or the Daily design, interval = day. For the Weekly design, interval = week.

PART 2: EVIDENCE TO SUPPORT THE VALIDITY OF ADAPTED SCALES

Adapting scales is not, in and of itself, a bad practice (Heggestad et al., 2019). However, adapting a scale necessarily raises concerns about the validity of the scores from the adapted scale (Aguinis & Vandenberg, 2014; American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014; Heggestad et al., 2019). Validity is the extent to which variation in standing on the latent construct causes variation in observed scores on a scale (Borsboom, Mellenbergh, & van Heerden, 2004). When some aspect of a scale is changed, then the strength of the relationship between the latent construct and the scores on the newly adapted scale will differ from the strength of the relationship between latent construct and the scores on the original scale. As such, it is imperative that when a scale is adapted the authors provide evidence to support the validity of the newly adapted scale, even when the adaptation is to a well-developed and well-validated scale.

In our sample of ESM studies, found that authors only infrequently provided evidence to support the validity of scales they adapted. Specifically, of the 364 scales authors reported as adapted, some form of validity evidence was provided in 70 (19.2%) of those cases. This failure to provide validation evidence is not specific to the ESM literature. For instance, in their review of the organizational science literature (not specific to ESM studies) Heggestad et al. (2019) found that authors provided evidence to support the validity of scales that they had adapted in 23% of such cases, a value quite similar to what we observed in our data.

When validity evidence for the adapted scale was reported within our sample, some form of factor analysis was the most common form of evidence provided: multilevel confirmatory factor analysis, confirmatory factor analysis, or exploratory factor analysis. Some authors, when

shortening a scale, examined part-whole correlations or used factor analytic information reported in the literature to select items with the highest factor loadings. Of course, the most appropriate form of evidence to support the validity of the adapted scale will depend on the nature of the adaptation. Heggestad et al. (2019) provide some guidance for the kinds of evidence that would be useful to support some of key forms of adaptations.

Citations in this Appendix

Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior, 1*(1), 569-595.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. 2004. The concept of validity. *Psychological Review,* 111: 1061–1071.

Heggestad, E. D., Rogelberg, S., Goh, A., & Oswald, F. (2015). Considering the effects of nonresponse on correlations between surveyed variables: A simulation study to provide context to evaluate survey results. *Journal of Personnel Psychology,* 14 (2), 91-103.

PART 3: RELIABILITY ESTIMATES OBSERVED IN OUR SAMPLE OF ESM STUDIES

In the paper, we noted that authors reported a point estimate of alpha 318 times, a range of values 95 times, and both a point estimate and a range 44 times. Across the 318 point estimates, the average reliability estimate is shown by scale length in Table 3.1. For example, across all 2-item scales for which a single, point-estimate of alpha was presented, the average alpha coefficient was .834. Across all three item scales, the average alpha was .854. Note that the average values of alpha shown in the table only increase slightly as scales get longer. The Spearman-Brown prophesy tells us that longer scales should tend to have higher levels of reliability, and the magnitude of the differences shown in Table 3.1 for longer scales are not as pronounced as would be expected. For example, taking the average alpha for scales of 8 or more items (i.e., .905), the Spearman-Brown prophesy suggests that the alpha for a 2-item scale should be .70, well below what we see in Table 6 (i.e., .834).

Table 3.1 also shows alpha values when a range was reported. Specifically, the table shows the means of the lowest and highest values by scale length. For example, across all 2-item scales the mean lowest value reported was .772 and the mean highest value reported was .872.

Table 3.1. Summary of reported Cronbach's alpha internal consistency reliability estimates

	Point Estimate *(n = 318)*	Range *(n = 95)*	
		Lower	Upper
2-item scales	0.834	0.772	0.872
3-item scales	0.854	0.761	0.873
4-item scales	0.865	0.824	0.911
5-item scales	0.865	0.753	0.863
6-item scales	0.864	0.846	0.899
7-item scales	0.905	0.915	0.965
8 or more items	0.905	0.854	0.928
Overall	0.866	0.810	0.898

Note: Values in the table are Cronbach's alpha internal consistency reliability coefficients. Both a point estimate and a range were presented for some scales.