Brose, Annette; Schmiedek, Florian; Gerstorf, Denis; Voelkle, Manuel C.

The measurement of within-person affect variation

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Emotion 20 (2020) 4, S. 677-699, 10.1037/emo0000583

The Measurement of Within-Person Affect Variation

Annette Brose[1,2,3], Florian Schmiedek[3,4], Denis Gerstorf[1], & Manuel C. Voelkle[1,3]

[1]Humboldt-Universität zu Berlin, Berlin, Germany

[2]KU Leuven, Leuven, Belgium

[3]Max Planck Institute for Human Development, Berlin, Germany

[4]German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany

Author Note

Correspondence concerning this article should be addressed to Annette Brose, Humboldt-Universität zu Berlin, Department of Psychology, Berlin, Germany. E-mail: annette.brose@hu-berlin.de

This manuscript was supported by a grant awarded to Annette Brose by the German Research Foundation [Deutsche Forschungsgemeinschaft, DFG], BR 3782/3-1. Denis Gerstorf also acknowledges the support provided by the German Research Foundation (DFG, GE 1896/6-1 and GE 1896/7-1). We thank Tim Grundmann for his work on the literature search that went into the review of this article, Elisabeth S. Blanke for comments on different versions of this manuscript, and two anonymous reviewers for excellent feedback on our work.

Abstract

The number of intensive longitudinal studies that investigate affective experiences at the within-person rather than the between-person level is rapidly increasing. This paradigmatic shift comes with new challenges such as questions revolving around how to measure within-person affect variation or more fundamental questions about the reliability and validity of constructs at the within-person level. We provide a review of substantive research published in *Emotion* since 2005, which revealed that to date no consensus has been established on measurement instruments for assessing within-person affective experiences. Our review also showed that researchers who are interested in within-person affect variation sometimes rely on measurement instruments that were established at the between-person level, which we think should be reconsidered. Finally, reliability estimates of state variation have been developed, but are not comprehensively reported in studies on within-person affect variation. The purpose of this paper is therefore to alert the reader to these issues and to highlight relevant criteria for selecting items and measurement instruments when studying within-person affect variation in intensive longitudinal studies. We recommend establishing common standards for measuring within-person affect variation, and to draw from a common pool of instruments because this would allow direct comparison of results across studies.

Word count abstract: 198

Word count main text: 7,917

Keywords: affect, within-person, reliability, intensive longitudinal studies, measurement

The Measurement of Within-Person Affect Variation

Inquiry into within-person affect variation has been a longstanding topic in the social and behavioral sciences (e.g., Lebo and Nesselroade, 1978). It has gathered considerable momentum over the past decade, with the number of studies that have collected numerous measurements within individuals over time rapidly increasing (for reviews, see Ong & Zautra, 2015; Röcke & Brose, 2013). Such studies commonly aim at understanding state variability or within-person dynamic phenomena such as emotion regulation, associations between stress and affect, or the structure of affect variation at the within-person level, to name just a few examples. This paradigmatic shift to within-person rather than between-person variation comes with an essential challenge: the question of how to measure affective experiences when the interest lies in variation within individuals across time. In this regard, several factors are important to consider: First, adjectives should be selected based on theoretical considerations and in view of the causes of variation at the within-person level (Schimmack, 2003; Brose, Voelkle, Lövdén, Lindenberger, & Schmiedek, 2015). Second, measurement instruments should be sensitive to within-person change and capture within-person variation reliably (e.g., Nezlek, 2016). Third, measurement instruments should be parsimonious to reduce participant burden in studies with intensive longitudinal designs (Cranford et al., 2006). While these remarks seem self-evident, we gained the impression that there is a lack of consensus on how to measure affective experiences at the within-person level. This impression is based on several observations from an exemplary review of studies in *Emotion* that have measured affect within individuals across time: (1) reasons are not made explicit for why items were selected and subscales composed to study within-person variation; (2) current choices of measurement instruments for the study of within-person variation are sometimes guided by questionnaires that have been established at the

between-person level, and these may have unknown psychometric properties at the within-person level; and (3) information on the reliability of within-person variation is often not (sufficiently) provided. These research practices come with several disadvantages, most importantly, with the difficulty to compare findings across studies and a lack of replicability that may in part be due to the diversity of research practices. The purposes of this article are therefore to allude to those common research practices, to discuss alternative approaches for selecting items and measurement instruments, and to discuss the reporting of reliability in studies on within-person variation of affective experiences.

Essential Concepts

Before we turn to the literature review, we elaborate on two essential concepts: The distinction of between-person vs. within-person variation in psychological research and the meaning of within-person reliability.

Between- vs. Within-Person Variation

For decades, research on affect was concerned with *between-person* variation in affect—the number and nature of the dimensions upon which persons differ, or the temperamental determination of between-person differences, to give two examples (Barrett & Feldman, 1998; Eysenck, 1970). Accordingly, measurement instruments were developed to capitalize on between-person variation. Figure 1.A illustrates such an instrument. Items assumed to be related to the same underlying construct—say, negative affect—differentiate well between three persons' levels of negative affect. Person 3 has the highest level of negative affect, and this is reflected in the scores of the specific items. Similarly, Persons 2 and 1 have item-specific scores in accordance with moderate and low levels of negative affect.

Meanwhile, when studying variation of affect at the *within-person* level, a measurement

instrument needs to capture affect variation well across occasions. See Figure 1.B for an illustration: Negative affect is lowest on the first occasion, and this is reflected in the scores of the three specific items. It is highest on Occasion 2 and moderate on Occasion 3, and again, the item-specific scores are in accordance with this variation across occasions.

Between- and Within-Person Reliability

If we were to estimate the reliabilities of the sets of items in Figure 1.A and 1.B, with a focus on their internal consistency (commonly referred to as Cronbach's Alpha in research on between-person variation; Cronbach, 1951; McDonald, 1999), we would find good reliabilities[1]. Please note the plural here: The variation of items across persons is consistent (Figure 1.A), and the variation of items across occasions is consistent (Figure 1.B). That is, there are two types of reliability for the different levels of analysis. Moreover, please note that the reliabilities are not perfect in these examples, as is common in psychological measurement. There are rank order changes of the items, but these are relatively small in comparison to the variation across persons (Figure 1.A) and the variation across occasions (Figure 1.B) and should thus be negligible at the construct level.

Reliability is generally defined as the ratio of systematic ("true") variance of interest to total observed variance. Reliability at the between-person level refers to the internal consistency of responses on one occasion across various persons on a set of items. Reliability at the within-person level refers to the internal consistency of responses within persons across occasions on a set of items (cf. Nezlek, 2017)—in colloquial terms, whether the ups and downs of items co-vary for these persons across time.

Current approaches to quantify within-person reliability are based on the idea of variance decomposition. Let's assume we measure anger with the items *angry*, *resentful*, and *annoyed* in a

study with *N* persons across *T* occasions, two of which are represented in Figure 1.C. Here, the total observed variance is due to (a) variation between persons (Person 2 exhibits higher levels of anger than Person 1), (b) variation within persons across occasions (anger is higher on Occasion 2 than on the other two occasions in both persons), (c) variation across items of the anger scale (on average the two persons seem to be more likely to describe themselves as angry rather than resentful), as well as due to (d) interactions of these sources and (e) measurement error. That is, systematic variance can be decomposed into multiple sources.

In the estimation of within-person reliability, systematic variance refers to the person × occasion interaction. This interaction reflects that a specific person has a unique true affect score that systematically varies across occasions, with this variation applying to all items of a given (sub)scale. The total variance relevant in the estimation of within-person reliability is the sum of this person × occasion interaction and the uniqueness of an observation (i.e., the random response error and the tendency of a person to provide a specific answer to a specific item on a specific occasion; Shrout & Lane, 2012).

In the following literature review, two specific estimation procedures were used for the estimation of within-person variability: estimation based on variance decomposition using generalizability analysis (referred to as GA-based in the following; Cranford et al., 2006) and estimation based on variance decomposition using multilevel modeling (referred to as MLM-based in the following; Bryk & Raudenbush, 1992; Nezlek, 2001, 2016). Both estimation procedures can be viewed as generalizations of Alpha in the generalizability theory (GT) framework. We will elaborate on these procedures below.

Together, affect can be measured with an interest in between-person and/or within-person variation. The reliability of measurement instruments in terms of internal consistency can also be

determined at both levels of analysis. Whereas measurement instruments and reliability estimation are well-established at the between-person level, the following review reveals a lack thereof at the within-person level.

Measuring Within-Person Affect Variation: A Review of Studies in *Emotion*

In our literature review, we searched for studies with intensive longitudinal designs that measured affective experiences across multiple occasions. In particular, we searched for articles that (i) were published in *Emotion* between 2005 and September 2017, (ii) used "experience sampling", "diary study", "ambulatory assessment", or "ecological momentary assessment" as key words, (iii) measured state affect on multiple occasions, and (iv) analyzed within-person variation of affective experiences. As can be obtained from Table 1, a total of 50 articles and 59 studies met these criteria. In the following, we will describe in detail how affect was measured in the referenced articles, with a focus on the selection of measurement instruments and items. We then turn to the issue of reliability estimates in those articles.

Selection of Measurement Instruments and Items

We have identified several approaches to the selection of measurement instruments and items in the reviewed studies (see Table 1, last column). The first approach has been to use items from established measures of between-person affect variation (22 studies; indicated by "based on bp measure"). The Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988), a well-established questionnaire for affect variation at the between-person level, was used in 10 of the 22 studies (3 studies used all PANAS items, 7 studies used different subsets of the PANAS items and some of those included additional items). Importantly, all studies that used PANAS items cited the work of Watson and colleagues who established the PANAS as a measure of between-person variation (e.g., Watson et al., 1988). In contrast, the PANAS has also

been investigated as a measure of within-person variation (Bleidorn & Peters, 2011; see below), but this study has not been cited in the reviewed studies. In the 22 studies, pooling of items was often in accordance with common pooling at the between-person level (e.g., aggregation of positive and negative affect items) or was not further detailed. One study analyzed the structure of within-person variation of the selected items (see Article 50) and justified pooling with these analyses. Together, the first approach indicates a reliance on established measurement instruments when investigating within-person affect variation, most importantly the PANAS. Yet, these instruments were established for the investigation of the between-person differences, and we will explain below why such reliance on between-person measures seems questionable.

As a second approach, eight studies used items that have previously been used in research on within-person variation (as indicated with "based on wp research" or "based on a wp measure" in the Details on Selection column). Of these studies, only two selected items in accordance with subscales that were previously shown to capture within-person affect variation reliably (see Article 5 and 20). Yet, the study reported in Article 20 pooled PA and NA items other than was reported in the referenced study. It pooled across high- and low-arousal positive and negative affective states. This contradicts the reasons provided for item selection (selection of high and low arousal states). The other studies did not provide information on the basis for pooling.

Studies that based the selection of items on existing measures (of between- or within-person variation) and on prior empirical research often did not mention conceptual or theoretical reasons for the selection of items (i.e., whether the study's interest pertained to a specific dimension of affect such as high arousal negative affect, or to discrete emotions such as fear or

anger; see Article 4 or 6).

As a third approach, the selection of items was based on theoretical accounts on emotions/affect in six studies (dimensional models, circumplex models, Lazarus's theory on emotions, discrete emotions, basis emotions). Albeit some reference to these accounts, reasons for specific selections of items from the often-large item pools were not provided in these studies. Article 3, for example, mentioned that item selection was based on the PANAS, but it only included four out of 20 items from the PANAS without further explanation. Moreover, in two studies, items were pooled in contradiction to the reasons provided for item selection (pooling across high and low arousal states; see Article Study 13 and 15). Four studies provided still other reasons for selecting items measuring affective experiences at the within-person level (e.g., own prior research), and a large number of studies (18) did not provide any information on the selection and pooling of items.

Together, affective experiences were measured very heterogeneously in the reviewed articles. In addition, there was an absence of clarity regarding the conceptual or analytic rationale for the selection of items as well as their aggregation in many studies. Indeed, the content of the emerging composite scores was not always clear as a result of the selection procedures (e.g., using subsets of PANAS items in combination with other items, using miscellaneous items from a circumplex model of affect). It remained unclear in multiple studies which dimensions, facets, content domains, or constructs of affect were under investigation, especially when the criteria for selecting items were not made explicit. This critique also includes our own reports (see Articles 1 and 4 in Table 1) just as it summarizes the emerging picture from the review. This observation is particularly striking in times during which a lack of replicability of findings across studies makes us reconsider the scientific practices in our field (Open Science Collaboration, 2015). Given the

variety of measures used, it remains an open question whether potential differences in findings across studies emerge from the different item sets that are used. Moreover, the lack of transparency of choices of particular measurement instruments or items over others seems to undermine the development of norms for the measurement of within-person affect variation because one cannot judge whether a specific item set has adequately fulfilled a given study purpose.

Reports of Reliability Estimates

We now describe how the reliability of the used measures and sets of items was determined and reported in the studies in our literature review. In 29 of the 50 articles, estimates of reliability were reported, which indicates some awareness of this issue (please see Columns 4 and 5 of Table 1). In the remaining 21 articles, reliability was either not reported (12 articles, containing a "not reported" in the Reliability column), or the analyses were done with single items (9 articles). Only 10 of the 29 articles that did report reliability estimates explicitly mentioned that reliability was estimated at the level of within-person variation (as indicated with "wp-var" in in the Estimation Procedure column). Seven of these articles reported within-person reliability estimates as suggested in the literature (as indicated with an asterisk; they reported MLM-based and GA-based estimates). In total, they reported the reliability estimates of 22 composites of items. In five cases, the estimates were between .5 and .6; in seven cases they were between .6 and .7; in four cases, they were between .7 and .8, and in six cases, they were higher than .8. The remaining studies that took a within-person approach to reliability estimation computed conventional Cronbach's Alpha or they computed the within-person correlation of items using the time series data of single individuals (see Articles 23, 34, and 46).

Of the remaining articles that did report reliability estimates, some reported estimates for

affect variation at the between-person level (three articles, as indicated with “bp-var” in the Estimation Procedure column). In these cases, item scores as observed on multiple occasions were pooled, and these pooled scores were the basis for reliability estimations. As Nezlek (2016) noted, such a reliability estimate would be similar to the reliability of trait-level measurement instruments if one considers means of states to be indicators of stable individual difference characteristics of persons/traits. Such between-person reliability estimates are not informative about within-person reliability. As another approach, reliability estimations in two of the reviewed studies were based on all data (i.e., the complete multilevel data set that contains information from multiple individuals on multiple occasions, as indicated by “mixed” in the Estimation Procedure column), thereby ignoring the multilevel structure of the data with occasions nested within persons. In these cases, it is not possible to determine whether reliability is based on systematic covariation between or within persons. In another 10 articles, no details were provided on how the reliability was estimated (as indicated by “no details” in the Estimation Procedure column). Since these studies simply reported “Alpha = “, sometimes with reference to Cronbach, we tend to think that these articles reported estimates of between- rather than within-person reliability. Finally, the type of data that went into the estimation of reliability remained ambiguous in four articles (as indicated by “unclear” in the Estimation Procedure column).

Together, the within-person reliabilities of the differently-composed sets of items in the studies reviewed in *Emotion* often remain unknown, in part due to inappropriate estimation procedures or the absence of reliability estimations. Established procedures for estimating within-person reliability were only used in seven of the reviewed articles. These articles revealed, however, that the within-person reliabilities of the item composites vary largely, with

some reporting reliability estimates that would have to be considered low using conventional standards for reliability at the between-person level. The diverse handling of reliability estimation in the reviewed studies, as well as the use of single items in the analyses of within-person variation, seems to indicate the necessity of a more elaborate understanding of this issue. The reliability of measurement instruments is one key determinant of whether associations with other constructs that exist can be found in empirical studies—at both the within-person level and the between-person level. Again, considering the challenges involved in replicability, one might question whether potential differences in findings across studies would be attributable in part to potentially low reliabilities of some of the item sets used. Given these insights from our literature review, we now elaborate on how the quality of measurement of within-person affect variation could be improved.

Improving the Quality of Measurement of Within-Person Affect Variation

This section has three parts in which we provide general recommendations on how to approach within-person affect variation in intensive longitudinal studies, with a view to guiding future decisions regarding study design and data analysis. In brief, the recommendations pertain to (1) the conceptual rationale for selecting specific facets of affect and measurement instruments; (2) the estimation of reliability at the within-person level; and (3) adequate parsimony of the measurement instruments.

Transparency of the Rationale for Selecting Measurement Instruments and Items

In the reviewed studies, information on the selection of items was sometimes limited to sentences like “affect was measured with the items x, y, and z”. When considering the diverse literature on affect which encompasses literature on mood and emotions, it is difficult to infer precisely what is meant by “affect” in this sentence without further information provided. To

mention just two conceptual frameworks of affect, the term is essential in the bipolar model of affect (Barrett & Russell, 1998) and in the broad distinction of the dimensions positive and negative affect by Watson and colleagues (1999). These frameworks clearly differ from one another in that the first postulates pleasure and activation as the two basic dimensions of affective experiences, whereas the second postulates the orthogonal dimensions positive and negative affect as the basic dimensions. It thus seems highly relevant to state precisely which theoretical framework underlies one's terminology. Moreover, it should be explicitly stated which particular aspects of a dimensional model of affect (e.g., "high arousal positive affect in accordance with the dimensional model of Watson et al.") or which discrete emotional states one wants to measure.

Relatedly, it is important to draw clear links between the theoretical framework, the specific construct, as well as the specific items under study. That is, one needs to consider the validity of the measure in light of the theoretical construct. In this context, it is important to consider the possibility that a construct qualitatively differs across the between-person and the within-person level of analysis (Geldhof et al., 2014). Some studies revealed structural differences in how affect varies across and within individuals. For example, Vansteelandt and colleagues (Vansteelandt, Van Mechelen, & Nezlek, 2005) modeled between-person and within-person affect variation and concluded that a dimensional model applies to emotional traits but not to emotional states. The latter are better characterized by a discrete and thus more differentiated model (see also Larsen & Zelinksi & Larsen, 2000). A more differentiated structure at the within-person level was also revealed by studies that analyzed the correlation between positive and negative affect. These found a more negative correlation at the within-person in comparison to the between-person level (e.g., Brose et al., 2015). Explanations for

these differences across levels are, in brief, that within-person variation is partly caused by phenomenologically distinct classes of situations, whereas between-person variation is partly determined by other, stable person-level characteristics (e.g., personality; Brose et al., 2015). These thoughts will be picked up in more detail in the context of reliability estimation and together with the use of the PANAS in within-person studies.

In sum, communicating and explicating in a transparent manner (i) the chosen theoretical framework of affect, (ii) the specific construct under study at the within-person level, and (iii) the link between items and the study purposes as well as the theoretical construct allow other researchers to make an informed evaluation of these choices. Indeed, if the choices seem adequate, this may motivate other researchers to follow them in their own research, thus contributing to more consistent methods across studies and more comparable findings.

Estimating the Reliability at the Within-Person Level

General recommendations. In view of how the reviewed articles dealt with the reliability issue, the first recommendation is that within-person reliabilities should always be estimated and reported, irrespective of whether a set of items was shown to be reliable in a previous study. A measurement instrument can be reliable in one study, but not reliable in another because within-person reliability is a function of a specific population of persons and a specific population of situations under study. Low reliability may occur, for example, if the variation of items across occasions or in a specific sample is low. For instance, if one measures anxious mood with the items *anxious*, *on edge*, and *uneasy* in students who take an exam (Cranford et al., 2006), the reliability is likely higher than if one measures anxious mood when the students of the same sample are on holidays. A reduction in reliability would be attributable to the likely reduction in variation in anxious mood. To give another example, older adults are

less variable in their experience of negative affect than younger adults (Röcke, Li, & Smith, 2009). If one measures negative affect in a study with an age-heterogeneous sample, it may thus occur that the reliability of the measure is good for younger adults and not good for older adults. The difference lies in the specifics of the two samples here, among them reduced variability in older adults.

Consequently, when choosing measurement instruments or items it is important to consider the population and the context and whether or which facets of affect are volatile for this population and in this context. Only if the context of some study leads to similar variation over time relative to another study and in the same population, reliability estimates can be expected to be comparable across studies. Related to the preceding, it is desirable that studies report information on variability in addition to reliability estimates (the average within-person standard deviation and the related standard deviation and range). This way, reliability estimates can be viewed in relation to variability, which is particularly useful when comparing potentially diverging reliability estimates of identical sets of items across studies. Relatedly, to get a sense of whether items that are commonly used in studies of between-person variation are sensitive to within-person variation, it is desirable to report the intraclass correlation (the proportion of between-person variation to total variation).

Improving within-person reliability. Figure 1.B illustrates a reliable within-person measure. The three items vary relatively consistently across time. Put differently, the measurement error is relatively small compared to the within-person true score variation (i.e., the within-person variation of negative affect). General ways to reduce measurement error are the following. First, one should formulate items as specific and unambiguously as possible. This includes choosing semantically unambiguous items and offering exact information on the time

scale the item refers to. Instead of asking "indicate how you feel", one can provide a specific time scale (e.g., "indicate how you have felt in the last five minutes"). Second, one can average across replicate measurement occasions of the same process. That is, if one is interested in day-to-day variations of mood, averaging across multiple within-day occasions to get a daily estimate increases the reliability in comparison to a single measurement of daily mood. Averaging replicate measures also means to average across items of parallel forms of measures (i.e., two replicate measures of the same construct). A third way to increase reliability is the choice of more homogeneous items. Inconsistency across items in within-person variation can have two sources: random error and non-homogeneity of the items (i.e., items that do not represent the construct of interest to the same degree). We will explain in the following that selecting more homogeneous items to increase reliability seems particularly appropriate at the within-person level. Please note that we will argue from the perspective of reflective measurement models. Such models assume the existence of latent variables / constructs (e.g., anger, intelligence). Furthermore, they assume that variation on observed variables that measure some construct is caused by variation of true scores at the construct level (Edwards & Bagozzi, 2000). As noted in the section on item selection, the causes of affect variation at the within-person level are often related to changes in the environment. Affect changes when one perceives changes in one's environment as pleasant or unpleasant, and specific appraisal processes determine whether one feels angry or anxious in a specific situation (e.g., being insulted may elicit feeling anger, whereas hearing about a disease may elicit anxiety). Figure 2.A illustrates this idea. It shows within-person variation in sadness, anger, and anxiety, each measured reliably with three items. Importantly, variation in the three different facets of affect is closely tied to the type of situation (as indicated on the *x*-axis of the graph): Sadness is highest in a loss situation, anger is highest

during conflict, and anxiety is highest when experiencing harm / threat. That is, the different situations elicit phenomenologically distinguishable feelings. Situations that cause one to feel angry are commonly not the same as those that cause people to feel sad or anxious. The extent to which affect variation is context-dependent, heterogeneity of items prevents consistent within-person covariation across all nine items represented in Figure 2.A. Yet, the homogeneity of the items at the subfacet level (three items each) should result in reliable measurements. If we were to choose to work with only one item from each of the sub-facets (e.g., sad, angry, and scared, see Figure 2.B for illustration) and estimate within-person reliability of such a composite across the three situations, the reliability estimate would be low. Again, the heterogeneity in content in interaction with situational characteristics prevents common covariation. Together, we think that it is more generally in accordance with the causes of variation at the within-person level to measure affective experiences in differentiated ways (Brose et al., 2015; Vansteelandt, Van Mechelen, & Nezlek, 2005; Zelenski & Larsen, 2000). To get reliable item composites, we propose measuring circumscribed facets of affect with homogeneous items rather than global dimensions.

Please note that the unique variances of items, which can reduce homogeneity and estimates of within-person reliability that are based on internal consistency, can nevertheless represent systematic sources of variability, so that the estimated reliability underestimates the true reliability of a composite score calculated from a set of items. Moreover, reliability estimates, as any estimate of a statistical coefficient that is based on a limited sample of persons and/or situations, are not free from error of estimation. Both implies that the reliability of within-person measures might be underestimated to some degree.

Distinguishing between higher-order and lower-order constructs. The preceding

reflections do not necessarily imply that affect measurement in intensive longitudinal studies is limited to a circumscribed scope of the affect space. It is of course possible to measure multiple, narrowly-defined facets (as is the case in Article #29 of the review or in the measure provided by Cranford et al., 2006 [see below]). As two items per facet are in principle enough to get good reliabilities at the within-person level and for circumscribed facets of affect (Wilhlem & Schoebi, 2007; Article 29 of the review), the scope of the affect space that is considered in a study can be extended without much additional burden for participants. Over and above this, once items are identified that measure narrowly-defined facets reliably, they could be combined in some theoretically justified way to create a higher-order construct. A recent investigation of within-person affect variation as observed with the day reconstruction method provides an example for the latter idea (Möwisch, Schmiedek, Richter, & Brose, 2018). This study investigated the structure of affect as measured with ten items. A series of multilevel confirmatory factor analyses revealed that a model with three subfacets of negative affect and a higher-order factor fit the data better than a model in which all affect items loaded on the same factor. That is, modeling separate lower-level subfacets and integrating those at a higher level was superior to modeling one latent affect factor with heterogeneous indicators.

Comparison of different classes of within-person reliability estimates.

We have mentioned above that two procedures were used for the estimation of within-person reliability estimation in the reviewed studies, estimation based on variance decomposition using generalizability analysis, and estimation based on variance decomposition using multilevel modeling. These fall into one class of within-person reliability estimation, namely the generalization of Alpha in the GT framework (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; see also Shrout & Lane, 2012). This class can be distinguished from a second class, the

generalization of Omega in latent variable models (the factor analytic tradition; Geldhof et al., 2014). Omega is a reliability index introduced in the factor analytic tradition, and it is conceived of as a more general reliability index with several advantages over alpha (McDonald, 1999; Raykov, 1997). We now elaborate on the differences between the two classes and then turn to recommendations on when to use which kind of estimate.

Cronbach's coefficient Alpha is a well-known measure of internal consistency of a set of items. It is defined as the ratio of systematic to total variance and was developed in the context of classical test theory (CTT). For the generalization of Alpha in the GT framework, it is essential that GT "recognizes multiple sources of variance in a given observed score" (Cranford et al., 2006, p.918) as well as their interactions. In the context of intensive longitudinal data, one essential source of variance is time (or measurement occasions), and the variance component of the interaction of person × time is used for the estimation of within-person reliability (see above). Despite this extension to multiple sources of systematic variance, essential features of the originally developed coefficient Alpha in CTT remain the same for reliability estimates in the GT framework: The item-construct relations are assumed to be homogeneous and error variances are assumed to be equal across items. That is, the items of a measurement instrument are considered parallel versions of each other, which means that the items are replicate indicators of the construct and thus are associated with the true score to the same degree (i.e., they are tau equivalent; Shrout & Lane, 2012).

The second class of within-person reliability estimates, the generalization of coefficient Omega, stands in the factor-analytic tradition of latent (reflective) variable models. The strength of the relationships between items and an underlying construct is allowed to vary across items– in factor-analytic terms, the items may have different loadings. Such models are also referred to

as congeneric measurement models. Omega is a reliability measure that is calculated on the basis of the loadings from these models. It thus allows for differential relations between individual items and their underlying construct, and the item-specific residual variances may vary. That is, item-construct relations are allowed to be heterogeneous in the case of Omega, and some items might reflect a construct to a stronger degree than others. In recent years, Omega has been generalized to the within-person level. Lane and Shrout (2011; see also Fuller-Tyszkiewicz et al., 2016) took a person-specific approach and presented how Omega can be obtained via dynamic factor analysis for single individuals. Geldhof et al. (2014) introduced the generalization of Omega in the context of multilevel confirmatory factor analysis (MLCFA).

The two classes of estimates have different advantages. Determining within-person reliability in the context of latent variable models (i.e., the generalization of Omega) might be particularly well-suited if one works with items that are heterogeneous in content because item-construct relations are not constrained to be equal. As was discussed in the context of research on between-person differences, this assumption is often violated, which is one reason that coefficient Alpha often underestimates the true reliability (Graham, 2006; Raykov, 1997). In cases in which items are rather homogeneous, estimating reliability in the GT framework will likely lead to similar results as estimation in the factor-analytic tradition.

A constraint of factor-based models is, however, that it requires a minimum of three indicators of a latent variable to compute composite reliability. To the contrary, it is an advantage of reliability estimation in the GT framework is that it can be accomplished even in the case of two indicators per construct.

When choosing the GT framework for within-person reliability estimation, one should keep another aspect in mind that has not been mentioned yet: that is, whether measurement

occasions are meaningfully ordered or not, or, put differently, whether studies have fixed or random designs. In fixed designs, data are collected on relatively consistent occasions across participants, and the ordering of occasions has some meaning (e.g., a series of measurement occasions leading up to an exam as is the case in the first study of Cranford et al., 2006; cf. Nezlek, 2016). In random designs, measurement occasions are not grouped across participants, and this is often the case in intensive longitudinal studies. Instead, occasions may be person-specific such as in event-contingent samplings (e.g., samplings are contingent on social interactions) or time-contingent samplings (e.g., when samplings are prompted by the passage of time). The GA-based estimation as described above should be used if study designs are fixed. In the case of random designs, reliability estimation is based on three nested variance components: items (nested within occasions), occasions (nested within persons), and persons (Nezlek, 2017; Nezlek & Gable, 2001). In accordance with the GT tradition, within-person reliability is the ratio of systematic to total variance here. The systematic variance components are the occasion-level variance component; the total variance component is the sum of occasion-level variance and variance across items.

Details on the different classes of within-person reliability estimation are provided in several valuable publications. Shrout and Lane (2012) recap basic ideas on CTT and reliability estimation before they discuss within-person reliability estimation in the GT framework and the factor-analytic tradition, together with illustrative examples, and SAS and SPSS syntax for reliability estimation in the GT framework. Cranford et al. (2006) introduce GA-based within-person reliability estimation for fixed designs. Nezlek (2017; cf. Nezlek & Gable, 2001) focus on within-person reliability estimation for random designs in the GT framework and provide HLM and Mplus syntax. A description of within-person reliability estimation in the factor-analytic

tradition, together with Mplus code, was provided by Geldhof et al. (2014). For methods in the GT framework, users of R can use the mlr function in the psych library for within-person reliability estimation.

Adequate Parsimony of Measurement Instruments

When measuring within-person affect variation in intensive longitudinal studies, one needs to keep the burden for participants in mind—they do not answer how they feel once, but do so multiple times. To give two examples of study protocols: there are diary studies with daily assessments on few occasions (e.g., eight occasions; Birditt, Fingerman, & Almeida, 2005) or experience sampling studies with many prompts across one day (every 15 minutes during the wakening hours; Ebner-Priemer & Sawitzki, 2007). Furthermore, one should keep the study purposes in mind. If the study aim is to thoroughly investigate affective experiences (e.g., on emotional complexity or linkages between emotion regulation and various specific affect facets in daily life), the range of measured facets of affect needs to be larger. If, however, affect is only one of multiple variables of interest, for example when one is interested in capturing as many aspects of daily life as possible, it may be enough to measure the valence dimension of affect, which should reflect the general ups and downs in daily life well. Hence, parsimony is advisable in many cases, but different study purposes might lead to diverging decisions on how many facets of affect to include in a study.

Established Measurement Instruments and Empirical Example

Having pointed at current research practices and ways to improve those, we now present studies on instruments specifically developed for measuring within-person affect variation. Furthermore, we summarize findings from a study that was concerned with the within-person reliability of the PANAS. We highlight those aspects of the studies that seem to be in accordance

with the recommendations just made. Moreover, we corroborate the illustrations with an empirical example.

We are aware of three studies that introduced measurement instruments of within-person affect variation. First, a brief 6-item measure designed to assess three dimensions of mood (Wilhelm & Schoebi, 2007) explicitly measures diffuse affective states, in contrast to specific emotions that are commonly elicited by specific events. It is based on a three-dimensional conceptualization of mood that distinguishes between the dimensions valence (ranging from unpleasant to pleasant), calmness (ranging from restless/tense to calm/relaxed), and energetic arousal (ranging from tired/low energy to awake/full of energy; see Steyer, Schwenkmezger, Notz, & Eid, 1997). Even though each dimension was measured with two bipolar items only, the within-person reliability of the three subscales (MLM-based estimates) was acceptable (.70 for valence and calmness, .77 for energetic arousal). With the given number of items per facet, it would not be possible to estimate within-person reliability in the factor analytic tradition / in analogy to Omega. In sum, this measurement instrument measures clearly defined aspects of mood, is reliable, and is very economic.

The second instrument is the shortened version of the Profile of Mood States (POMS; Cranford et al., 2006), originally a measure of between-person differences (McNair, Lorr, M, & Droppleman, 1992). It consists of 15 items and taps into five facets of affect, each indexed by three items: anxious mood, depressed mood, anger, vigor, and feelings of fatigue. The within-person reliabilities of the subscales were acceptable to good (.75 to .88, GA-based estimates) in two samples and different contexts, with the exception of the subscale “depressed mood” in one of the two samples (.62). The shortened POMS does not clearly distinguish between discrete feelings (e.g., anger) and aspects of mood (e.g., fatigue). Yet, it is noteworthy that it captures

within-person affect variation in a broad range of affective experiences (five facets) with acceptable reliability using merely 15 items. In addition, the selection of the subscales was, from a content-oriented perspective, in line with the particular purposes of the studies, which were the examination of affective experiences during a stressful period in peoples' lives and in intimate relationships (Bolger, Zuckerman, & Kessler, 2000; Kennedy, Bolger, & Shrout, 2002). With three items per facet, it would also be possible to apply congeneric measurement models here and estimate within-person Omegas for each subscale.

A third study recently investigated the dimensionality of affect variation within- and between individuals in a sample of children (Leonhardt et al., 2016). Affect was measured with 20 items in this study. Multilevel CFA revealed the existence of three correlated mood dimensions in the daily lives of children (good–bad mood, alertness–tiredness, and calmness–tension). Given that the articles identified in our literature review were concerned with within-person affect variation in adults, we will not go into more detail on this measurement instrument.

Next to these developments of new instruments, another study examined the within-person reliability of the PANAS (Bleidorn & Peters, 2011). The PANAS is known to have good reliabilities at the between-person level (Leue & Beauducel, 2011). This study tested the reliability (MLM-based estimates) of the two PANAS subscales (PA and NA, 10 items each) with two different data sets. Reliability estimates were good for the PA subscale (> .85) but not acceptable for the NA subscale (.63 and .50, respectively). Thus, the PANAS did not measure negative affect reliably here, despite the use of more items than were used per facet by Cranford and colleagues (2006) and Wilhelm and Schoebi (2007). Notably, reliability estimates would likely be higher if the authors had modeled congeneric measurement models.

Bleidorn and Peters (2011) did not elaborate on a potential match between the PANAS

subscales' items and true score affect variation at the within-person level. To this end let us consider potential causes of affect variation as measured with the PANAS NA subscale as well as the diverse content of the NA items (e.g., distressed, nervous, guilty, and hostile). Neuroticism is assumed to be one cause of between-person variation in NA, and this trait results in enhanced levels of diverse aspects of affect (i.e., people with high levels of neuroticism tend to experience more distress, nervousness, guilt, and hostility than people with low levels of neuroticism). In contrast, distinct causes of within-person affect variation are commonly linked to specific changes in the environment (see above; Figure 1.D), leading to more nuanced affective reactions. Distress, nervousness, guilt, and hostility therefore likely co-vary less at the within-person than at the between-person level. As a consequence of the level-specific causes of variation, we would have expected the reliability of the PANAS NA subscale to be high at the between-person level, but not necessarily at the within-person level. At the within-person level, item variation does not seem to reflect variation of the same latent variable. Put differently, the scope of the NA subscale is broad and this might be the reason why the reliability of the NA subscale of the PANAS is questionable at the within-person level.

To corroborate these observations as well as the above recommendations, we now provide an empirical example. In the COGITO study, 101 younger and 103 older adults visited a laboratory on about 100 occasions (M = 101, range = 87 – 109 occasions) to work on cognitive tasks and fill out an electronic diary (Schmiedek, Lövdén, & Lindenberger, 2010). At the beginning of each session, participants evaluated their current affect on the PANAS items. Table 2 provides within-person reliability estimates for (a) the whole NA subscale; (b) different subsets of NA items as used in some of the studies reviewed in Table 1; and (c) three pairs of items from the PANAS that belong to three different content categories (Zevon & Tellegen, 1982). We focus

on the NA subscale for demonstration purposes. As for the reliability estimates, we provide estimates in the GT framework for random designs and estimates in the factor-analytic tradition (i.e., within-person generalizations of Alpha and Omega).

As expected, within-person reliability estimates from the GT class were in all cases smaller than the factor-based estimates. That is, not allowing for individual item-construct relations as is the case in the estimation of Alpha seems to underestimate reliability. Relating this to Bleidorn et al. (2011), the quite low reliabilities as reported for the PANAS NA subscale would likely be higher if they had chosen a different approach. Another obvious finding is that the reliability estimates were always smaller in the older subsample. One reason for this sample specificity might be the age group differences in variance. As reported in the supplement, older adults varied less on all items in comparison to younger adults (average η^2 = .28, range .20 to .37).

Turning to the different subscale compositions, there are several noteworthy aspects. When viewing the GA-based estimates, subscales with more items are generally more reliable, which is in accordance with the Spearman-Brown formula. However, this general pattern is not ubiquitous. In particular, the subscales "fearful" (especially when applied to younger adults) and "guilty" (which respectively only have two items), are more reliable in this study than the subscales with three, five or six items when comparing them with the other estimates in the GT framework. In line with our reasoning above, our explanation for this pattern is that the causes of variation at the within-person level affect the respective items of the "fearful" and "guilty" subscales rather equally, and thus lead to a pattern of strong covariation. Put differently, the homogeneity of the items in these subscales fits well with the specificity of situation-driven within-person affect variation, as also is the case in the subscales examined by Crandford et al.

(2006; see above). Instead, the five-item composite by Murray et al. (2009, Table 1), for example, samples items from five different content categories (distressed, fearful, jittery, angry, and guilty) as listed by Zevon and Tellegen (1982). Here, it is unclear which series of person x situation interactions could possibly drive covariation among these items.

Of further note, viewing the reliability of the content category "fearful" in the older adults together with information on the variability at the item level, it is obvious that low variability as present in older adults' experiences of fear does not necessarily undermine reliability. Instead, it seems that the two similar indicators of "fearful" occur very consistently across time, even though their frequency of occurrence and/or intensity of variation is low. The low reliability of the content category "distressed" may have emerged because the indicators do not seem to belong to the same content domain, at least in the German translation of the PANAS that was used in this study. One indicator directly reflects distress, whereas the second, "verärgert", reflects anger rather than distress.

Together, by providing estimates for two age groups, this example demonstrates the relevance of specific sample characteristics for reliability estimates. It also speaks for the importance of considering the causes of affect variation at the within-person level. Furthermore, it is a case in which within-person estimates from the GT class are lower than estimates in the factor-analytic tradition, probably speaking for the underestimation of reliability in the former.

Discussion

The interest in within-person psychological processes has grown rapidly over the last decade, and many studies measuring within-person affect variation have been conducted. Here, we exemplarily reviewed such studies that were published in *Emotion* between January 2005 and September 2017. From our selective literature review, we concluded that no consensus has been

established to date regarding how to measure within-person affective variation. Moreover, we have observed some reliance on established measurement instruments of between-person affect variation and noted shortcomings in the provision of within-person reliability estimates. With this backdrop, we provided recommendations on how to improve the quality of the measurement of within-person affect variation. We emphasized (1) that the theoretical and analytic rationale for selecting and measuring specific facets of affect should be made transparent; (2) which aspects to keep in mind for improving reliability; and (3) that measurement instruments should be (adequately) parsimonious. These recommendations were exemplified by pointing at publications that promote economic and reliable within-person measurement instruments of affect. In this context, we also provided an example in which the use of an established measure of between-person variation, the PANAS, did not lead to satisfying results regarding the reliability estimates of within-person affect variation. We attributed one essential difference between the instruments to the scope of measures. Those with narrowly defined facets were superior to the PANAS that samples heterogeneous items. These observations are in accordance with those from the literature review. Those measurement instruments with circumscribed content were generally shown to have reasonable reliabilities (e.g., Article 29 in Table 1), whereas measurement instruments that in contrast included states from clearly different content domains lacked acceptable reliability (e.g., Article 20 in Table 1).

A notable limitation of this study's literature review is that its generalizability is narrower than the relevance of our recommendations. We focused on intensive longitudinal studies that used diary or experience sampling methodologies. That is, other types of studies with instances of measuring within-person affect variation were not considered. This includes experimental research (e.g., emotional responding to filmclips; Kunzmann & Grühn, 2005), longitudinal

studies on emotional development (e.g., following critical life events; Luhmann, Hofmann, Eid, & Lucas, 2012), or research on clinical interventions in which affective experiences are often an important outcome variable (e.g., Gunthert, Cohen, Butler, & Beck, 2005). We thus cannot comment on whether research practices in these areas meet established standards more consistently than in the reviewed literature.

A second limitation of our work is that we cannot make clear recommendations on how to treat items with low variability. As we pointed out in the context of the example, while low variability may constrain reliability, this does not necessarily have to be the case. Furthermore, states with low frequency and / or variation in intensity may be relevant to a construct and to some criterion variable, which speaks against their exclusion from research on within-person variability in daily life. To date, there is a lack of systematic methodological research on the relationship between within-person variability and reliability, and on how non-normally distributed variables affect within-person reliability estimates. Thus, for the time being it is a matter of investigators deciding within each empirical study on the relevance of items with low variability, and piloting the properties of selected items.

Our focus on narrowly defined facets of affect, in the service of improving reliability, warrants some reflection. We argued from the perspective of reflective measurement models (see above) and claimed the importance of within-person reliability and the consideration of level-specific causes of variation. As a consequence, we recommended the use of measurement instruments with a circumscribed scope (cf. Cronbach & Gleser, 1965, for discussion of the bandwith-fidelity dilemma). This proposition may be qualified in different ways:

First, it seems important to not constrain the debate on methodological aspects of how to measure within-person affect variation entirely on the aspect of reliability in terms of internal

consistency. Think of the example of failure at school for which negative feelings may be an important predictor. With regard to the outcome (failure at school), it may not be relevant whether a student is sad or angry. What matters is the perturbation of feelings independent of the particular facet of negative affect, and it might also be relevant to sample infrequent states here (i.e., items with low variability such as shame or guilt; please see Table 1 of the supplement). Thinking of the causes of variation in this example, it would perhaps be appropriate if the items were pooled in accordance with a formative measurement model. In formative measurement models, constructs are conceived of as composites of specific component variables, that is, the construct is formed by its measures (Edwards & Bagozzi, 2000). The measures are commonly, yet not necessarily, correlated, but their correlation does not have to be based on a common cause (i.e., the variation of construct). At school, perturbations of specific feelings (sadness-, anger-, shame- or guilt-related feelings as in the example) may have different causes. For example, sadness may occur in situations of social rejection and anger may occur in situations of social conflict. The causes of variation thus differ. Still, regarding the predictive validity of negative affect for failure at school, it would be advisable to consider heterogeneous items irrespective of their causes of variation and frequency of occurrence when computing a composite score. In other words, measuring anger, sadness, guilt, or shame opposed to measuring either of them may not reflect a common cause of variation well. Yet, it is relevant to capture variation in both because failure at school depends on both.

Second, the use of overly homogeneous items at the within-person level could prevent insights into more or less nuanced within-person responding. In the case of affective experiences, individuals are known to show more or less differentiated responses to events (Erbas, Ceulemans, Blanke, Sels, Fischer, & Kuppens, in press). Whereas some people feel either sad or

disappointed in a specific situation, others have a more global response in that they feel mainly good or bad, and a third group of people may be able to feel two aspects of affect simultaneously (e.g., sadness and disappointment). These individual differences in affect differentiation are related to more global aspects of emotional functioning and personality. A prerequisite for the study of nuanced affective responding is the inclusion of items that are somewhat diverse in content—if items are too homogeneous, individual differences in how nuanced people react to events could not be investigated.

A final important issue that we would like to raise is that we argued from the perspective of sample-based research and analyses and hence provided reliability estimates that hold at the average level. That is, we describe how affect varies within "average" individuals (e.g., context-specific variation, Figure 1.D) and how reliability is estimated if within-person variation is pooled across participants. Yet, variation and co-variation of affect across time varies across individuals (Brose et al., 2015), and both affects reliability estimates. A recent study analyzed within-person reliability (Omega) of affect at the individual level using dynamic factor analysis (Fuller-Tyszkiewicz et al., 2016). It reported large individual differences in reliability estimates, ranging from 0.18 to 1.00, while the average reliabilities of the positive and negative affect subscales were acceptable. Importantly, individual differences in reliability had little impact on fixed effects in multilevel models (i.e., sample-based estimates) as subsequent analyses revealed. However, the results imply that if the goal of the research is to examine individual persons, then reliability should be studied at the individual level.

A further possible approach for dealing with individual differences in within-person processes has been proposed by Nesselroade and colleagues: The proposition of the idiographic filter that allows idiosyncratic measurement models (Nesselroade, Gerstorf, Hardy, & Ram,

2007). The idea behind this proposition is that the general construct, for example negative affect, is invariant across individuals while the construct-observable relations (i.e., loadings in factor models) are not. It allows affect perturbation to have a different specific meaning for different individuals. Think once more of affect perturbation at school. It could reflect the experience of anger in one student (one with externalizing behavioral tendencies) and sadness in another (one with internalizing behavioral tendencies). Yet, one would expect affect perturbation to have similar relationships with other constructs (e.g., a positive association with failure at school, independent of the relative importance of its indicators in different students).

Together, the aspects to consider in studies on within-person affect variation go beyond the recommendations we have made in this manuscript. However, these aspects seem to be special cases that exist side by side with the more general principles that we have formulated. The selection of measurement instruments and items should be based on criteria that require theoretical, methodological, and statistical considerations. These criteria should also be kept in mind when adopting well-established measurement instruments of between-person variation for the study of within-person variation. We also propose that nothing speaks against the composition of new sets of items along the lines of clear criteria. In the long run, it is of course desirable that researchers draw from a pool of shared and established measurement instruments because this would allow direct comparison of results across studies. To date, such standards are not available in research on within-person affect variation. We hope our work provides insights that in the long run help establish such standards and thereby increase the interpretability and replicability of findings across studies.

References

Barrett, L. F., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology, 74,* 967–984.

Becker, E. S., Keller, M. M., Goetz, T., Frenzel, A. C., & Taxer, J. L. (2015). Antecedents of teachers' emotions in the classroom: an intraindividual approach. *Frontiers in Psychology, 6*: 635. doi: 10.3389/Fpsyg.2015.00635

Birdett, K. S., Fingerman, K. L., & Almeida, D.M. (2005). Age differences in exposure and reactions to interpersonal tensions: A daily diary study. *Psychology and Aging, 20*, 330-340.

*Blanke, E. S., Riediger, M., & Brose, A. (2018). Pathways to happiness are multidirectional: Associations between state mindfulness and everyday affective experience. *Emotion, 18,* 202-211. doi: 10.1037/emo0000323

Bleidorn, W., & Peters, A. L. (2011). A multilevel multitrait-multimethod analysis of self- and peer-reported daily affective experiences. *European Journal of Personality, 25*, 398-408. doi: 10.1002/Per.804

Bolger, N., Zuckerman, A., & Kessler, R. C. (2000). Invisible support and adjustment to stress. *Journal of Personality and Social Psychology, 79*, 953-961. doi: 10.1037//0022-3514.79.6.953

*Brans, K., Koval, P., Verduyn, P., Lim, Y. L., & Kuppens, P. (2013). The regulation of negative and positive affect in daily life. *Emotion, 13*, 926-939. doi: 10.1037/a0032400

*Bresin, K., Fetterman, A. K., & Robinson, M. D. (2012). Motor control accuracy: A consequential probe of individual differences in emotion regulation. *Emotion, 12*, 479-

486. doi: 10.1037/a0025865

*Brose, A., Lövdén, M., & Schmiedek, F. (2014). Daily fluctuations in positive affect positively co-vary with working memory performance. *Emotion, 14*, 1-6. doi: 10.1037/a0035210

Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (2015). Differences in the between-person and within-person structures of affect are a matter of degree. *European Journal of Personality, 29*, 55-71. doi: 10.1002/per.1961

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods* (1st ed.). Newbury Park, CA: Sage.

*Cameron, L. D., & Overall, N. C. (2017). Suppression and expression as distinct emotion-regulation processes in daily interactions: Longitudinal and meta-analyses. *Emotion*, Advance online publication. doi: 10.1037/emo0000334

*Catalino, L. I., Arenander, J., Epel, E., & Puterman, E. (2017). Trait acceptance predicts fewer daily negative emotions through less stressor-related rumination. *Emotion*, Advance online publication. doi: 10.1037/emo0000279

*Chin, A., Markey, A., Bhargava, S., Kassam, K. S., & Loewenstein, G. (2017). Bored in the USA: Experience sampling and boredom in everyday life. *Emotion, 17*, 359-368. doi: 10.1037/emo0000232

*Chue, A. E., Gunthert, K. C., Ahrens, A. H., & Skalina, L. M. (2017). How does social anger expression predict later depression symptoms? It depends on how often one is angry. *Emotion, 17*, 6-10. doi: 10.1037/emo0000239

*Compton, R. J., Arnstein, D., Freedman, G., Dainer-Best, J., Liss, A., & Robinson, M. D. (2011). Neural and behavioral measures of error-related cognitive control predict daily

coping with stress. *Emotion, 11*, 379-390. doi: 10.1037/a0021776

*Conner, T., & Barrett, L. F. (2005). Implicit self-attitudes predict spontaneous affect in daily life. *Emotion, 5*, 476-488. doi: 10.1037/1528-3542.5.4.476

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi: 10.1007/BF02310555

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

*De Leersnyder, J., Koval, P., Kuppens, P., & Mesquita, B. (2017). Emotions and concerns: Situational evidence for their systematic co-occurrence. *Emotion*, Advance online publication. doi: 10.1037/emo0000314

*Denissen, J. J. A., Butalid, L., Penke, L., & van Aken, M. A. G. (2008). The effects of weather on daily mood: A multilevel approach. *Emotion, 8*, 662-667. doi: 10.1037/a0013497

Ebner-Priemer, U. W., & Sawitzki, G. (2007). Ambulatory assessment of affective instability in borderline personality disorder-the effect of the sampling frequency. *European Journal of Psychological Assessment, 23*, 238–247. doi: 10.1027/1015-5759.23.4.238

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of the relationship between constructs and measures. *Psychological Methods, 5*, 155-174.

Erbas, Y., Ceulemans, E., Blanke, E., Sels, L., Fischer, A. H., & Kuppens, P. (in press). Emotion

differentiation dissected: Between-category, within-category, and integral emotion differentiation, and their relation to well-being. *Cognition and Emotion.*

Eysenck, H. J. (1970). *The structure of human personality*. London: Methuen.

Fuller-Tyszkiewicz, M., Hartley-Clark, L., Cummins, R.A., Tomyn, A., Weinberg, M., & Richardson, B. (2016). Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. *Psychological Assessment, 29*, 1120-1128.

*Genet, J. J., & Siemer, M. (2012). Rumination moderates the effects of daily events on negative mood: Results from a diary study. *Emotion, 12*, 1329-1339. doi: 10.1037/a0028070

Gunthert, K. C., Cohen, L. H., Butler, A., & Beck, J. (2005). Predictive role of daily coping and affective reactivity in cognitive therapy outcome: Application of a daily process design to psychotherapy research. *Behavior Therapy, 36*, 77-88.

*Heiy, J. E., & Cheavens, J. S. (2014). Back to basics: A naturalistic assessment of the experience and regulation of emotion. *Emotion, 14*, 878-891. doi: 10.1037/a0037231

*Hill, C. L. M., & Updegraff, J. A. (2012). Mindfulness and its relationship to emotional regulation. *Emotion, 12*, 81-90. doi: 10.1037/a0026355

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*, 72-91. doi: 10.1037/a0032138

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. What they are and how to use them. *Educational and Psychological Measurement, 66*, 930-944. doi: 10.1177/0013164406288165.

*Hoorelbeke, K., Koster, E. H., Demeyer, I., Loeys, T., & Vanderhasselt, M.-A. (2016). Effects of cognitive control training on the dynamics of (mal)adaptive emotion regulation in

daily life. *Emotion, 16*, 945-956. doi: 10.1037/emo0000169

Huang, P. H., & Weng, L. J. (2012). Estimating the reliability of aggregated and within-person centered scores in ecological momentary assessment. *Multivariate Behavioral Research, 47*, 421-441. doi: 10.1080/00273171.2012.673924

*Iijima, Y., Takano, K., & Tanno, Y. (2017). Attentional bias and its association with anxious mood dynamics. *Emotion*, Advance online publication. doi: 10.1037/emo0000338

*Jiang, D., Fung, H. H., Sims, T., Tsai, J. L., & Zhang, F. (2016). Limited time perspective increases the value of calm. *Emotion, 16*, 52-62. doi: 10.1037/emo0000094

*Juslin, P. N., Liljeström, S., Västfjäll, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: Listener, music, and situation. *Emotion, 8*, 668-683. doi: 10.1037/a0013505

*Kashdan, T. B., & Farmer, A. S. (2014). Differentiating emotions across contexts: Comparing adults with and without social anxiety disorder using random, social interaction, and daily experience sampling. *Emotion, 14*, 629-638. doi: 10.1037/a0035796

*Keng, S.-L., & Tong, E. M. (2016). Riding the tide of emotions with mindfulness: Mindfulness, affect dynamics, and the mediating role of coping. *Emotion, 16*, 706-718. doi: 10.1037/emo0000165

Kennedy, J. K., Bolger, N., & Shrout, P. E. (2002). Witnessing interparental psychological aggression in childhood: Implications for daily conflict in adult intimate relationships. *Journal of Personality, 70*, 1051-1077.

*Klipker, K., Wrzus, C., Rauers, A., & Riediger, M. (2017). Hedonic orientation moderates the association between cognitive control and affect reactivity to daily hassles in adolescent

boys. *Emotion, 17*, 497-508. doi: 10.1037/emo0000241

*Koval, P., Brose, A., Pe, M. L., Houben, M., Erbas, Y., Champagne, D., & Kuppens, P. (2015). Emotional inertia and external events: The roles of exposure, reactivity, and recovery. *Emotion, 15*, 625-636. doi: 10.1037/emo0000059

*Koval, P., & Kuppens, P. (2012). Changing emotion dynamics: Individual differences in the effect of anticipatory social stress on emotional inertia. *Emotion, 12*, 256-267. doi: 10.1037/a0024756

*Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion, 13*, 1132-1141. doi: 10.1037/a0033579

Kunzmann, U. & Grühn, D. (2005). Age differences in emotional reactivity: The sample case of sadness. *Psychology and Aging, 20*, 47-59. doi: 10.1037/0882-7974.20.1.47

Leue, A. & Beauducel, A. (2011). The PANAS structure revisited: On the validity of a bifactor model in community and forensic samples. *Psychological Assessment, 23*, 215-225.

Lebo, M. A., & Nesselroade, J. R. (1978). Intraindividual differences dimensions of mood change during pregnancy identified in five P–technique factor analyses. *Journal of Research in Personality*, *12*, 205–224.

Leonhardt, A., Könen, A., Dirk, J., & Schmiedek, F. (2016). How differentiated do children experience affect? An investigation of the within-and between-person structure of children's affect. *Psychological Assessment, 28*, 575-585. doi: 10.1037/pas0000195

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4*, 192-211. doi:

10.1037/1082-989x.4.2.192

Luhmann, M., Hofmann, W., Eid, M., & Lucas, R. E. (2012). Subjective well-being and adaptation to life events: A meta-analysis. *Journal of Personality and Social Psychology, 102*, 592–615. doi: 10.1037/a0025948

*Luong, G., Wrzus, C., Wagner, G. G., & Riediger, M. (2016). When bad moods may not be so bad: Valuing negative affect is associated with weakened affect-health links. *Emotion, 16*, 387-401. doi: 10.1037/emo0000132

McNair, D. M., Lorr, M., & Droppleman, L. F. (1992). *EdITS manual for the Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Service.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

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

Möwisch, D., Schmiedek, F., Richter, D., & Brose, A. (2018). Capturing affective well-being in daily life with the day reconstruction method: A refined view on positive and negative affect. *Journal of Happiness Studies*. Advance online publication.

*Morelli, S. A., Lee, I. A., Arnn, M. E., Zaki, J., & Morelli, S. A. (2015). Emotional and instrumental support provision interact to predict well-being. *Emotion, 15*, 484-493.

*Murray, G., Nicholas, C. L., Kleiman, J., Dwyer, R., Carrington, M. J., Allen, N. B., & Trinder, J. (2009). Nature's clocks and human mood: The circadian system modulates reward motivation. *Emotion, 9*, 705-716. doi: 10.1037/a0017080

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

Nesselroade, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Idiographic filters for

psychological constructs. *Measurement: Interdisciplinary Research and Perspectives, 5*, 217-235. doi: 10.1080/15366360701741807

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

Nezlek, J. B., & Gable, S. L. (2001). Depression as a moderator of relationships between positive daily events and day-to-day psychological adjustment. *Personality and Social Psychology Bulletin, 27*, 1692-1704. doi: 10.1177/01461672012712012

*Nezlek, J. B., Vansteelandt, K., Van Mechelen, I., & Kuppens, P. (2008). Appraisal-emotion relationships in daily life. *Emotion, 8*, 145-150. doi: 10.1037/1528-3542.8.1.145

*O'Hara, R. E., Armeli, S., Boynton, M. H., & Tennen, H. (2014). Emotional stress-reactivity and positive affect among college students: The role of depression history. *Emotion, 14*, 193-202. doi: 10.1037/a0034217

*Ode, S., Hilmert, C. J., Zielke, D. J., & Robinson, M. D. (2010). Neuroticism's importance in understanding the daily life correlates of heart rate variability. *Emotion, 10*, 536-543. doi: 10.1037/a0018698

Ong, A. D., & Zautra, A. J. (2015). Intraindividual variability in mood and mood regulation in adulthood. In M. Diehl, K. Hooker, & M. Sliwinski (Eds.), *Handbook of Intraindividual Variability Across the Lifespan* (198-215). New York, NY: Routledge.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*: 6251. doi: 10.1126/science.aac4716

*Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., & Nicolson, N. A. (2006). Diurnal mood variation in major depressive disorder. *Emotion, 6*, 383-391. doi: 10.1037/1528-

3542.6.3.383

*Pond, R. S., Jr., Kashdan, T. B., DeWall, C. N., Savostyanova, A., Lambert, N. M., & Fincham, F. D. (2012). Emotion differentiation moderates aggressive tendencies in angry people: A daily diary analysis. *Emotion, 12*, 326-337. doi: 10.1037/a0025762

Rauers, A., Blanke, E., & Riediger, M. (2013). Everyday empathic accuracy in younger and older couples: Do you need to see your partner to know his or her feelings? *Psychological Science, 24*, 2210-2217. doi: 10.1177/0956797613490747

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329-353.

*Riediger, M., Wrzus, C., Schmiedek, F., Wagner, G. G., & Lindenberger, U. (2011). Is seeking bad mood cognitively demanding? Contra-hedonic orientation and working-memory capacity in everyday life. *Emotion, 11*, 656-665. doi: 10.1037/a0022756

*Riediger, M., Wrzus, C., & Wagner, G. G. (2014). Happiness is pleasant, or is it? Implicit representations of affect valence are associated with contrahedonic motivation and mixed affect in daily life. *Emotion, 14*, 950-961. doi: 10.1037/a0037711, 10.1037/a0037711.supp (Supplemental)

*Righetti, F., Gere, J., Hofmann, W., Visserman, M. L., & Van Lange, P. A. (2016). The burden of empathy: Partners' responses to divergence of interests in daily life. *Emotion, 16*, 684-690. doi: 10.1037/emo0000163

*Robinson, M. D., Moeller, S. K., Buchholz, M. M., Boyd, R. L., & Troop-Gordon, W. (2012). The regulatory benefits of high levels of affect perception accuracy: A process analysis of

reactions to stressors in daily life. *Emotion, 12*, 785-795. doi: 10.1037/a0029044

Röcke, C., & Brose, A. (2013). Intraindividual variability and stability of affect and well-being: Short-term and long-term change and stabilization processes. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry, 26,* 185–199. doi:10.1024/1662-9647/a000094

Röcke, C., Li, S.-C., & Smith, J. (2009). Intraindividual variability in positive and negative affect over 45 days: Do older adults fluctuate less than young adults? *Psychology and Aging, 24*, 863-878. doi: 10.1037/a0016276

Schimmack, U. (2003). Affect measurement in Experience Sampling research. *Journal of Happiness Studies, 4*, 79-106. doi: 10.1023/A:1023661322862

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience, 2*, 1-10. doi: 10.3389/fnagi.2010.00027

*Shackman, A. J., Weinstein, J. S., Hudja, S. N., Bloomer, C. D., Barstead, M. G., Fox, A. S., & Lemay, E. P., Jr. (2017). Dispositional negativity in the wild: Social environment governs momentary emotional experience. *Emotion*, Advance online publication. doi: 10.1037/emo0000339

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302-320). New York: Guilford.

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen. Handanweisung [The Multidimensional Mood Questionnaire*

(MDMQ)]. Göttingen, Germany: Hogrefe.

*Stone, A. A., Schwartz, J. E., Schkade, D., Schwarz, N., Krueger, A., & Kahneman, D. (2006). A population approach to the study of emotion: Diurnal rhythms of a working day examined with the day reconstruction method. *Emotion, 6,* 139-149. doi: 10.1037/1528-3542.6.1.139

*Takano, K., Sakamoto, S., & Tanno, Y. (2013). Ruminative self-focus in daily life: Associations with daily activities and depressive symptoms. *Emotion, 13*, 657-667. doi: 10.1037/a0031867

*Takano, K., & Tanno, Y. (2011). Diurnal variation in rumination. *Emotion, 11*, 1046-1058. doi: 10.1037/a0022757

*Thompson, R. J., Kuppens, P., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Gotlib, I. H. (2015). Emotional clarity as a function of neuroticism and major depressive disorder. *Emotion, 15*, 615-624. doi: 10.1037/emo0000067

*Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Gotlib, I. H. (2011). Concurrent and prospective relations between attention to emotion and affect intensity: An experience sampling study. *Emotion, 11*, 1489-1494. doi: 10.1037/a0022822

*Tong, E. M. W., Bishop, G. D., Enkelmann, H. C., Why, Y. P., Diong, S. M., Khader, M., & Ang, J. (2005). The use of ecological momentary assessment to test appraisal theories of emotion. *Emotion, 5*, 508-512. doi: 10.1037/1528-3542.5.4.508

*Tong, E. M., & Jia, L. (2017). Positive emotion, appraisal, and the role of appraisal overlap in positive emotion co-occurrence. *Emotion, 17*, 40-54. doi: 10.1037/emo0000203

*van Roekel, E., Verhagen, M., Engels, R. C., & Kuppens, P. (2017). Variation in the serotonin transporter polymorphism (5-HTTLPR) and inertia of negative and positive emotions in

daily life. *Emotion*, Advance online publication. doi: 10.1037/emo0000336

Vansteelandt, K., Van Mechelen, I., & Nezlek, J. B. (2005). The co-occurrence of emotions in daily life: A multilevel approach. *Journal of Research in Personality, 39*, 325-335.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070.

*Westgate, E. C., Wilson, T. D., & Gilbert, D. T. (2017). With a little help for our thoughts: Making it easier to think for pleasure. *Emotion, 17*, 828-839. doi: 10.1037/emo0000278

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life. Structural validity, sensitivity to change, and reliability of a short scale to measure three basic dimensions of mood *European Journal of Psychological Assessment, 23*, 258-267.

*Wong, E., Tschan, F., & Semmer, N. K. (2017). Effort in emotion work and well-being: The role of goal attainment. *Emotion, 17*, 67-77. doi: 10.1037/emo0000196

*Wrzus, C., Luong, G., Wagner, G. G., & Riediger, M. (2015). Can't get it out of my head: Age differences in affective responsiveness vary with preoccupation and elapsed time after daily hassles. *Emotion, 15*, 257-269. doi: 10.1037/emo0000019

*Wrzus, C., Wagner, G. G., & Riediger, M. (2014). Feeling good when sleeping in? Day-to-day associations between sleep duration and affective well-being differ from youth to old age. *Emotion, 14*, 624-628. doi: 10.1037/a0035349, 10.1037/a0035349.supp (Supplemental)

Zelenski, J. M., & Larsen, R. J. (2000). The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data. *Journal of Research in Personality, 34*, 178-197.

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic

analysis. *Journal of Personality and Social Psychology, 43*, 111-122.

Footnotes

[1] When speaking of reliability in this manuscript, we refer to internal consistency if not noted otherwise. Estimates of reliability assume that all observed variables measure a single latent variable.

Table 1.

Affect Measurement in Intensive Longitudinal Studies, Emotion January 2005 – September 2017.

	Article	Positive Affective States Items and Subscales	Negative Affective States Items and Subscales	Reliability of Subscales	Estimation Procedure	Details on Selection / Pooling of Items
1	Blanke et al., 2017	happy, relaxed, content	distressed, nervous, downhearted	PA: .76 NA: .64	GA-based* *wp-var*	"structure of affect was confirmed using multilevel confirmatory factor analyses in Mplus" Selection: / Pooling: *empirical justification*
2	Brans et al., 2013	happy, relaxed	angry, stressed, anxious, depressed	Studies 1 & 2 PA: .58/.65 NA: .65/.66	MLM-based* *wp-var*	not provided
3	Bresin et al., 2012	positive affect: enthusiastic, excited empathy: caring, empathetic	negative affect: distressed, nervous anger: annoyed, angry	range: .64 - .78	α, *no details*	"items were generally selected on the basis of Watson and Clark's (1994) psychometric work", empathy items "on the basis of Meier et al. (2006)" Selection: *based on bp measures, PANAS, vague* Pooling: /
4	Brose et al., 2014	all PANAS items	distressed, upset, irritated, nervous, jittery	not reported	n.a.	PANAS (Watson, Clark, & Tellegen, 1988), aggregation of PA and NA subscale Selection: *based on bp measure, PANAS* Pooling: *based on bp research*
5	Cameron et al., 2017		depressed mood: sad, lonely, hopeless, discouraged fatigue: worn out, exhausted	Studies 1 - 4 depressed mood: .83-.89 fatigue: .69-.83	α, *no details*	depressed mood: Cranford et al., 2006 fatigue: McNair, Lorr, & Droppleman, 1992 Selection: *based on wp and bp measures, POMS* Pooling: *based on wp and bp research*
6	Catalino et al., 2017		disgust, guilt, shame, anger, contempt, embarrassment, hatred, sadness, fear, and anxiety	.83	α, *no details*	modified Differential Emotion Scale (Fredrickson, Tugade, Waugh, & Larkin, 2003); aggregation across all NA items Selection: *based on bp measure* Pooling: /
7	Chin et al., 2017	excitement, interestedness, alertness, confidence, love, contentedness, happiness, hopefulness, relief	anger, boredom, exhaustion, frustration, indifference, loneliness, sadness, overwhelmingness, worry	n.a. *analyses of single items*	n.a.	"emotion measures were selected by the firm administering the study, and we remain agnostic as to whether these measures reflect specific emotions as characterized by academic research" Selection: *arbitrary* Pooling: *n.a.*

8	Chue et al., 2017		angry, sad, anxious	not reported	n.a.	"Similar items have been used successfully in a number of daily process studies (Forand, Gunthert, German, & Wenze, 2010; Wenze, Gunthert, & Forand, 2010)"; Selection: *previous wp research, vague* Pooling: *wp reliabilities for item composite not provided in referenced studies*
9	Conner & Barret, 2005	active, alert, interested, proud, enthusiastic, aroused, surprised, peppy, joy, happy, amused, satisfied, calm, relaxed, quiet, still	ashamed, guilt, nervous, afraid sleepy, sluggish, tired, bored, sad, disappointed, disgust, embarrassed, angry	n.a. *analyses of single items*	n.a.	"Affect terms represented all combinations of valence (pleasant– unpleasant) and arousal (high–low activation) dimensions of the affective circumplex (Feldman, 1995)" Selection: *previous wp research, circumplex, vague* Pooling: n.a.
10	Compton et al., 2011	all PANAS items	all PANAS items	not reported	n.a.	PANAS; aggregation of PA and NA subscale Selection: *based on bp measure, PANAS* Pooling: *based on bp research*
11	De Leersnyder et al., 2017	<u>positive disengaging (1)</u>: proud about myself, elated/ exuberant, happy/joyful <u>positive engaging (2)</u>: relying on another, close to another	<u>negative disengaging (3)</u>: angry, disappointed, contemptuous, sad <u>negative engaging (4)</u>: ashamed, indebted	(1): .69 (2): .80 (3): .80 (4): .63	α, *no details*	"adapted version of Emotional Patterns Questionnaire", "specific emotion terms were selected because of their high factor loadings in the previous self-report studies"; Selection: *based on bp measure* Pooling: *referenced studies investigated bp structure*
12	Denissen et al., 2008	all PANAS items	all PANAS items <u>tiredness</u>: sleepy, tired, sluggish, drowsy, quiet, still	not reported	n.a.	"Daily positive and negative affect were assessed by means of the PANAS mood scale (Watson & Clark, 1994)"; items "from PANAS-fatigue scale (Watson & Clark, 1994) loaded on the same factor as the items "quiet" and "still" that tap into the arousal dimension of the mood circumplex (Feldman Barrett, 1995), so they were combined into a single scale of daily tiredness" Selection: *based on bp measure, PANAS* Pooling: *based on bp analyses*

13	Genet & Siemer, 2012		high arousal: upset, afraid, guilty, nervous, ashamed, angry, disgusted, embarrassed low arousal: sad, tired, depressed, down, annoyed, worried	.91	α, "across days and participants" *mixed /bp-var*	"Affect items were selected to measure negative affect high and low in arousal in the affective circumplex model of Feldman Barrett and Russell (1998)" Selection: *circumplex, high and low arousal* Pooling: *aggregation into one NA score irrespective of level of arousal*
14	Heiy & Cheavans, 2014		unpleasant mood; if present: evaluation of content: anger, anxiety/fear, embarrassment/ shame, guilt, disgust, sadness, loneliness	n.a. *analyses of single items*	n.a.	not provided
15	Hill & Updegraff, 2012	interested, proud, happy, content, peaceful, calm, overjoyed, fascinated, curious, comfortable	ashamed, nervous, irritated, afraid, guilty, sad, angry, enraged, depressed, miserable, fearful	PA: .90 NA: .89	α, "alpha for mean" PA and NA *unclear*	"Emotions varied on the dimension of pleasantness–unpleasantness and were representative of both high and low activation emotions" Selection: *dimensional model: valence and arousal* Pooling: *pooling into one PA and NA score, irrespective of level of arousal*
16	Hoorelbeke et al., 2017	energetic, satisfied, happy	angry, tense, depressed	not reported	n.a.	Items were "adopted from the Profile Of Mood States (McNair, Lorr, & Dropplemann, 1992) in line with Rossi and Pourtois (2012)", aggregation of PA and NA subscale Selection: *based on wp and bp research, POMS* Pooling: *wp reliabilities not provided in referenced studies*
17	Iijima et al., 2017		five items	not reported	n.a.	"adjectives adopted from the tension-anxiety scale of the Profile of Mood States-Brief version (McNair, Lorr, & Droppleman, 1992)" Selection and pooling: *based on bp measure, POMS*
18	Jiang et al., 2016	high arousal: enthusiastic low arousal: calm		n.a. *analyses of single items*	n.a.	"…conceptually and empirically distinguished between actual and ideal high arousal positive affective states" used the "the Affect Valuation Index (AVI; Tsai et al., 2006) " Selection: *based on bp measure* Pooling: *n.a.*

19	Juslin et al. 2008	calm-contentment, happiness-elation, surprise-astonishment, interest-expectancy, love-tenderness, pleasure-enjoyment	shame-guilt, disgust-contempt, nostalgia-longing, anger-irritation, boredom-indifference, anxiety-fear, sadness-melancholia	n.a. *analyses of single items*	n.a.	emotion terms selected include the "basic" emotions typical of discrete emotion theories, such as anger, surprise, interest, and fear (Izard, 1977), cover all four quadrants of the "Circumplex" model in terms of valence and arousal (Russell, 1980), and feature typical music-related terms such as pleasure, nostalgia, and expectancy (Juslin & Laukka, 2004, Table 4)" Selection: *basic emotions and four quadrants of circumplex, music-related states* Pooling: *n.a.*
20	Kashdan & MacKnight, 2013	high arousal: enthusiastic, joyful low arousal: content, relaxed	high arousal: anxious/nervous, angry low arousal: sad, sluggish	PA: .64 NA: .59	MLM-based* *wp-var*	Distinguished between high and low arousal items "(Nezlek, 2005)" Selection: *based on wp research, high and low arousal* Pooling: *pooling into one PA and NA score, irrespective of level of arousal*
21	Keng & Tong, 2016	pride, amusement, contentment, hope, gratitude, joy, love, serenity	guilt, shame, anger, disgust, fear, frustration, sadness	PA: .92 NA: .96	α, average scores across time *bp-var*	not provided
22	Klipker et al., 2017	enthusiastic, content, happy, relaxed	angry, disappointed, sad, stressed	not reported	n.a.	not provided
23	Koval & Kuppens, 2012		threat emotion: anxious, stressed	not reported	n.a. / *wp-var*	aggregation based on size of MLM wp correlation Selection: not provided Pooling: *empirical justification*
24	Koval et al., 2013	happy, relaxed	sad, depressed, anxious, angry	not reported	n.a.	not provided
25	Koval et al., 2015		angry, sad, anxious, depressed	.71	MLM-based* *wp-var*	not provided
26	Luong et al., 2016	enthusiastic, interested, joyful, content, relaxed, well, energetic	nervous, angry, tired, downcast, disappointed, tense	not reported	n.a.	not provided, aggregation of PA and NA subscale

27	Morelli et al., 2015	excited, happy, joyful, elated	anxiety: upset, scared, anxious, stressed	PA: .81 NA: .88	α, *no details*	assessed anxiety and happiness as reported in Gable, Gosnell, Maisel, & Strachman, 2012 Selection and pooling: *based on bp-research*
28	Murray et al., 2009	Study 1: excited, active interested, determined Studies 2-3: all PANAS items Studies 1-3: happy	Study 1: upset, guilty, scared, hostile, jittery Studies 2-3: all PANAS items Studies 1-3: sad	Study 1 PA: .77 (*M*) NA: .64 (*M*)	α, *estimated for each occasion, bp-var*	Study 1: "each day, brief measures of PA and NA were created by abbreviating the well validated [PANAS] (Watson & Clark, 1997; Watson et al., 1988)", valence Studies 2-3: PANAS and valence Selection: *based on bp research, PANAS* Pooling: *based on bp research*
29	Nezlek et al., 2008	joy: content, happy love: sympathy, affection	guilt: ashamed, guilty fear: nervous, fear anger: irritation, angry sadness: sorrow, sad	joy: .97, love: .86, guilt: .58, fear: .51, anger: .95, sadness: .92	MLM-based* *wp-var*	"Following Lazarus (1991) we selected two positive and four negative emotions" Selection: *following Lazarus, vague* Pooling: *following Lazarus*
30	Ode et al., 2010		irritated, annoyed, dejected, sad	.89	α, "across days" *unclear*	Items "were taken from the [PANAS-X] (Watson & Clark, 1994), and we sought to include a balance of high- and low-arousal items" Selection: *based on bp research, PANAS, plus low arousal* Pooling: *aggregation into one NA score, irrespective of level of arousal*
31	O'Hara et al., 2014	enthusiastic, happy, content, cheerful	anxious affect: nervous, anxious hostile affect: hostile, angry depressed affect: sad, unhappy, dejected	PA: .89 depressed: .83 anxious: .72 hostile: .76	α, "across all person-days" *unclear*	"face valid items were chosen by the researchers based on Larsen and Diener's (1992) circumplex model of emotion" Selection: *circumplex, vague* Pooling: /
32	Peeters et al., 2006	enthusiastic, strong, happy, cheerful, talkative, satisfied, and self-assured	guilty, irritable, anxious, restless, tense, easily distracted, agitated	PA: .95 NA: .91	α, across all data *mixed /bp-var*	"… relied on items that were used previously by our research group in different populations"; aggregation based on results from PCA with aggregated scores (*bp level*) Selection: *previous own research* Pooling: *based on bp analyses*

33	Pond et al., 2012		Study 1 anger experience: angry, frustrated, provoked, hostile aggression: items not provided Study 2 anger: angry aggression: "three items that assessed their aggressive behaviors when provoked"	Study 1 anger: .91 aggression: .66 Study 2 aggression: .81	MLM-based* *wp-var*	Study 1: "intensity of anger experience" and "an abbreviated form of the physical […] and verbal […] aggression subscales of the Aggression Questionnaire" Study 2: intensity of anger, daily aggression Selection and pooling: *based on bp research*
34	Riediger et al., 2011	interested, joyful, content	angry, downcast, anxious	PA: .65 (*M*) NA: .52 (*M*)	α, *estimated for each individual, wp-var*	"The items were selected because they represent prototypical positive and negative affect facets that are relevant for, and evince sufficient intraindividual variation in, the daily lives of individuals from different age groups." Selection: *based on age-comparative wp research, vague* Pooling: /
35	Riediger et al., 2014	interested, enthusiastic, joyful, content, relaxed, energetic, balanced	angry, downcast, anxious, disappointed, tense, tired	not reported	n.a.	"The items were selected to represent prototypical positive and negative affect facets of various arousal levels that are relevant for, and evince sufficient intraindividual variation in, the daily lives of individuals from different age groups." Selection: *based on age-comparative wp research, vague* Pooling: /
36	Righetti et al, 2016	"I am in a positive mood"	"I am in a negative mood" stressed	negative and positive (reverse-coded) mood: .84	α; *no details*	not provided
37	Robinson et al., 2012		dejected, depressed	.88	α *no details*	"two common markers of depressive feeling states (Watson, 2000)" Selection & pooling: *prior research*
38	Shackman et al., 2017	cheerful, happy, joyful	nervous, anxious, worried	NA, PA > .96	α, *no details*	not provided

39	Stone et al., 2006	happy, warm/friendly, enjoying myself	hostile/angry, frustrated/annoyed, worried/anxious, criticized/put down, depressed/ blue, pushed around/ hassled	n.a. *analyses of single items*	n.a.	"adjectives are very similar to those used in other mood adjective checklists such as the Nowlis (Nowlis, 1965), POMS (McNair, Lorr & Droppelman, 1972), and PANAS (Clark, Watson & Leeka, 1989; Watson & Tellegen, 1985)" Selection: *based on bp measure, PANAS, POMS* Pooling: *n.a.*
40	Takano & Tanno, 2011		scared, afraid, upset, nervous, jittery, distressed	.88	α, *no details*	"negative mood states were assessed by six adjectives […] selected from the negative affect subscale of the [PANAS] ([…] Watson, Clark, & Tellegen, 1988)" Selection and pooling: *based on bp measure, PANAS*
41	Takano et al., 2013	active, proud, strong	scared, afraid, upset	PA, NA: > .99	α, *no details*	"Negative and positive mood states were each assessed using three adjectives […] selected from the [PANAS] (Watson, Clark, & Tellegen, 1988)" Selection and pooling: *based on bp measure, PANAS*
42	Thompson et al., 2011	active, alert, exited, happy	angry, ashamed, frustrated, guilty, sad, disgusted, anxious	PA: .82 NA: .81	α, "across experience sampling period" *unclear*	"affect words were drawn from various sources, including the [PANAS] (Watson, Clark, & Tellegen, 1988) and Ekman's basic emotions (Ekman, Friesen, & Ellsworth, 1972)" Selection: *based on bp measure, PANAS, and basic emotions* Pooling: /
43	Thompson et al., 2015	excited, happy	angry, sad, anxious	Study 1, PA: .72, NA: .66 Study 2: PA: .75, NA: .57	GA-based* *wp-var*	not provided
44	Tong et al., 2005	happiness		n.a. *analysis of single item*	n.a.	not provided

45	Tong et al., 2017	interest, pride, amusement, awe, challenge, contentment, hope, gratitude, joy, love, relief, serenity		n.a. *analyses of single items*		focus on these items because "[this] builds on and extends" prior work, "some of these emotions […] are featured in classic works by early [emotion] theorists", "they reflect different adaptation functions" Selection: *prior work, vague* Pooling: *n.a.*
46	Van Roekel et al., 2017	joyful, satisfied, happy, energetic, relaxed, cheerful	irritated, guilty, anxious, worried, low, insecure	NA: .69 PA: .70	α, wp *wp-var*	not provided
47	Westgate et al., 2017	valence dimension of mood, single item		n.a. *analyses of single items*	n.a.	not provided
48	Wong et al., 2017	interest, pride, enthusiasm, happiness, joy, pleasure, tenderness, relief, compassion	shame, guilt, anger, contempt, disgust, disappointment, anxiety, sadness, embarrassment	not reported	n.a.	Based on Geneva Emotion Wheel, "a self-report instrument of emotion developed by Scherer (2005) [that] asks about a variety of discrete emotions" Selection: *based on bp-research, discrete emotions* Pooling: /
49	Wrzus et al., 2014	enthusiastic, happy, energetic, even-tempered, content, relaxed	nervous, tense, angry, tired, downcast, disappointed	PA: .88 NA: .79	α, average across days, *bp-var*	not provided
50	Wrzus et al., 2015		activating: nervous, tense, angry deactivating: disappointing, tired, downcast	not reported	n.a.	"selected adjectives were from validated adjective lists to assess affect (Hampel, 1977; Matthews et al., 1990; Watson & Clark, 1999)", used MLCFA to test fit of the two-factor structure Selection: *based on bp-measures, PANAS* Pooling: *empirical justification*

Note. Comments of the authors are written in italics; PA = positive affect, NA = negative affect; * = estimation of within-person reliability in accordance with established procedures; GA-based = estimation based on generalizability theory analysis; MLM-based = estimation based on multilevel modeling; n.a. = not applicable; wp-var = within-person variation; bp-var = between-person variation; MLCFA = multilevel confirmatory factor analysis.

Table 2.

Reliability of the PANAS NA Subscale and Subsets of Various PANAS NA Items.

	Total NA subscale	Subset 6 items	Subset 5 items	Subset 5 items	Subset 3 items	Selection of 2 items each, distinction between content categories (Zevon & Tellegen, 1982)		
Reference Table 1		Takano & Tanno, 2011	Brose et al., 2012	Murray et al., 2009	Takano et al., 2013	Distinction between content categories (Zevon & Tellegen, 1982)		
Items	distressed, upset, irritable, hostile, afraid, scared, nervous, jittery, guilty, ashamed	distresed, upset, afraid, scared, nervous, jittery	distressed, upset, irritable, nervous, jittery	upset, guilty, scared, hostile, jittery	upset, afraid, scared	Fearful: afraid, scared	Guilty: guilty, ashamed	Distressed: distressed, upset
Younger adults								
Within-person reliability, GT based	0,76	0,66	0,67	0,53	0,53	0,73	0,68	0,52
Within-person reliability, factor based	0,84	0,77	0,75	0,7	0,63	/	/	/
Older adults								
Within-person reliability, GT based	0,41	0,30	0,36	0,00	0,35	0,67	0,42	0,22
Within-person reliability, factor based	0,75	0,7	0,69	0,56	0,51	/	/	/

Notes. Reliability in the factor analytic tradition can only be estimated for subscales with more than two items.

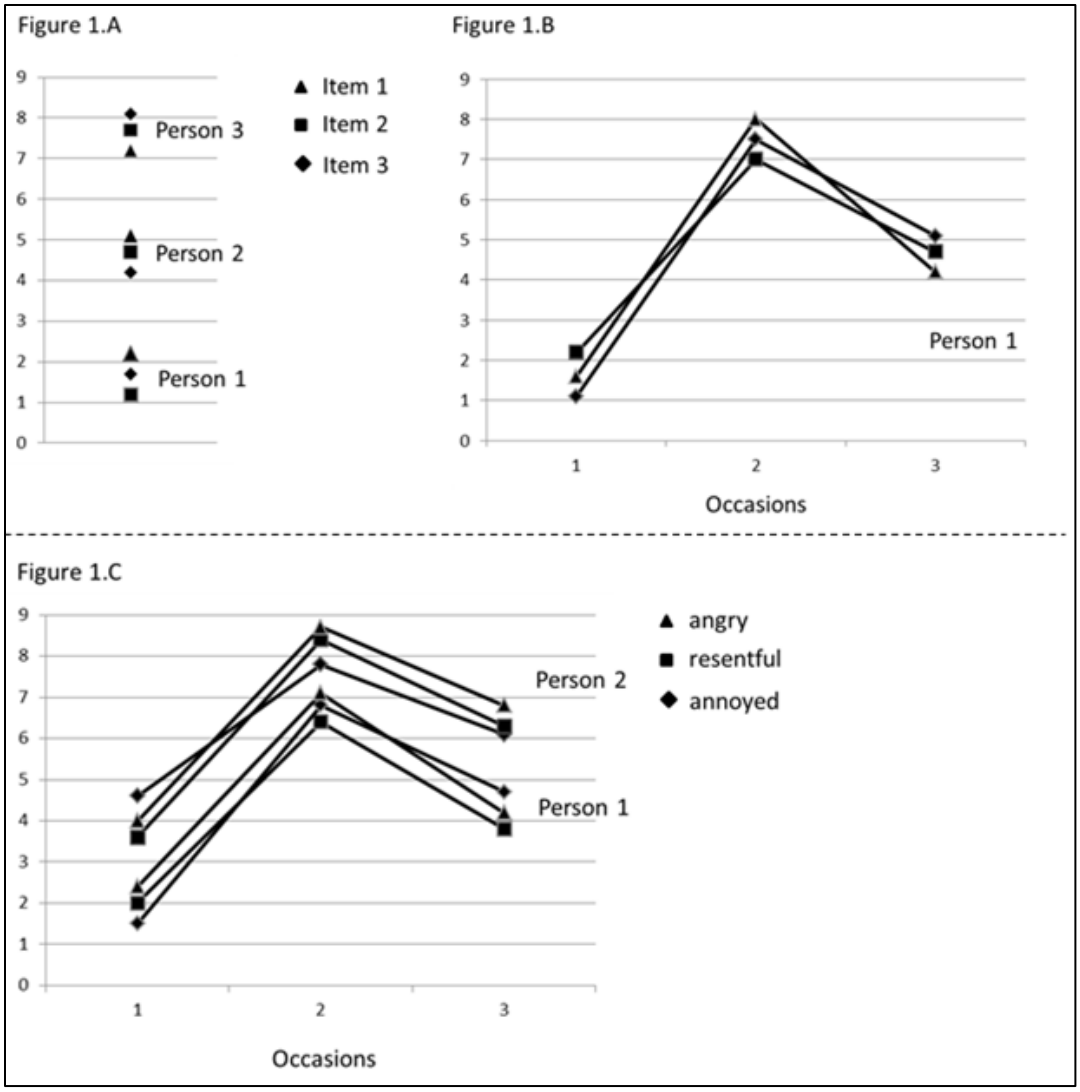

Figure 1. Illustration of reliable between-person (1.A) and within-person (1.B, 1.C) covariation. 1.A illustrates between-person variation of negative affect, 1.B illustrates within-person variation of negative affect, both measured with three items; 1.C illustrates within-person variation of anger, measured with three items. The examples are generated such that the internal consistencies would be good, but not perfect—as is common in psychological measurement. That is, there are rank order changes of the items, but these should be negligible at the construct level (i.e., they are relatively small in comparison to variation across person and / or occasions).

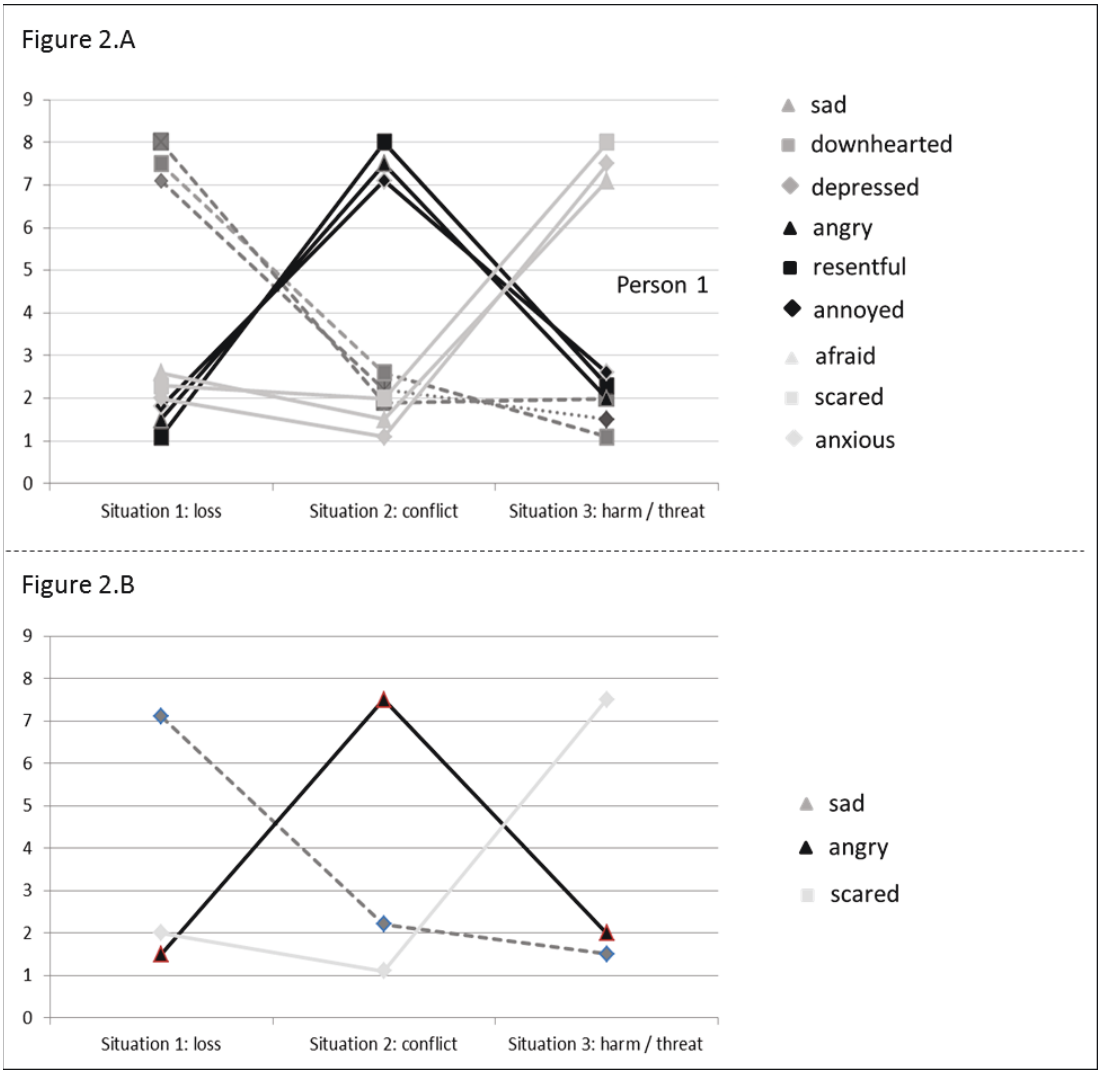

Figure 2. Figure 2.A illustrates within-person variation in three content categories, sadness, anger, and fear, each measured with three items and across three different types of situations. Figure 2.B illustrates within-person variation in three items from three different content categories: sadness, anger, and fear.