Original Research Article

Assessment
1–10
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10731911221113563
journals.sagepub.com/home/asm
SAGE

Examining the Concurrent and Predictive Validity of Single Items in Ecological Momentary Assessments

Jiyoung Song[1], Esther Howe[1], Joshua R. Oltmanns[2], and Aaron J. Fisher[1]

Abstract

Although single items can save time and burden in psychology research, concerns about their reliability have made the use of multiple-item measures the default standard practice. Although single items cannot demonstrate internal reliability, their criterion validity can be compared with multiple-item measures. Using ecological momentary assessment data, we evaluated repeated measures correlations and constructed multilevel cross-lagged models to assess concurrent and predictive validity of single- and multiple-item measures. Correlations between the single- and multiple-item measures ranged from .24 to .61. In 27 of 29 unique single-item predictor models, single items demonstrated significant predictive validity, and in one of eight sets of comparisons, a single-item predictor exhibited a larger effect size than its multiple-item counterpart. Although multiple-item measures generally performed better than single items, the added benefit of multiple items was modest in most cases. The present data provide support for the use of single-item measures in intensive longitudinal designs.

Keywords

single items, ecological momentary assessment, predictive validity, psychometrics, participant burden

Researchers have long argued that multiple-item measures are psychometrically superior to single-item measures, and the use of multiple-item scales is currently the default standard practice in psychological research (Clark & Watson, 2016; Viswanathan, 2005). When multiple items have a high average intercorrelation, often summarized with Cronbach's alpha, they are said to exhibit high internal reliability. Proponents thus believe that multiple-item measures are inherently more reliable than their single-item counterparts, which cannot demonstrate reliability in a similar manner (Churchill, 1979; Peter, 1979). Multiple-item measures also tend to provide more information than single-item measures. Reducing a potentially complex construct into a single question could fail to accurately capture the multifaceted target construct (Jacoby, 1978), whereas asking multiple questions about a construct may be able to detect relatively fine-grained differences in responses (Nunnally & Bernstein, 1994). Furthermore, multiple-item measures are thought to be less susceptible to sources of measurement error than single-item measures because latent variable models can disentangle shared and unique variance in item responses and even simple sum scores can average unintended noise in single-item measures (DeVellis, 2003).

The field's emphasis on the reliability of multiple-item measures, however, is not without challenges and concerns. First, the reliability argument is null for a singular and concrete target construct that can be captured by a single item (Rossiter, 2002), and Cronbach's alpha should be examined only after a multiple-item measure has shown to constitute a unidimensional construct (Cortina, 1993). The use of single items may be especially appropriate for unidimensional constructs with a sufficiently narrow meaning (Sackett & Larson, 1990). For example, depression is one of many highly heterogeneous conditions of which symptoms, such as sad mood, insomnia, concentration problems, and suicidal ideation, significantly differ from one another in their etiology and impact on impairment (Fried & Nesse, 2015). Depression thus could be considered to comprise a wide range of concrete and singular symptoms. Although a single item assessing one of the depressive symptom criteria is insufficient for measuring the diagnostic construct, it may still provide enough information for a given symptom of interest.

Furthermore, multiple-item measures suffer from common method variance. Responding accurately to every

[1]University of California, Berkeley, USA
[2]Stony Brook University, NY, USA

Corresponding Author:
Jiyoung Song, Department of Psychology, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94720, USA.
Email: jiyoungsong@berkeley.edu

question in a long survey demands substantial cognitive effort from participants (Krosnick, 1999). Potential fatigue may make participants less thorough, leading to stereotyped or uniform responses within and across measures (Viswanathan & Kayande, 2012). Altogether, these issues can result in altered values of multiple-item measures and, subsequently, potentially inflated intercorrelation.

Solutions for assessing the reliability of single items have been offered. Wanous and Hudy (2001) provide two potential approaches for assessing the reliability of a single item. First, these authors propose solving for the single-item reliability within the correction for attenuation formula (Nunnally & Bernstein, 1994). A complete discussion of this approach and its assumptions is outside the scope of the present study; however, it is worth emphasizing that Wanous and Hudy (2001) suggest that the correlation between the single item and its scale form be held at 1.00 in pursuing this solution. Given that such an assumption undermines the goals of the present study—namely, to evaluate the relative performance of a single item and a comparison scale—we do not employ this method here. Alternatively, these authors note that because the total variance in a factor model is the sum of the communality, specificity, and unreliability, communality can be used as a conservative estimate of reliability for a single item in the absence of specific variance (Weiss, 1976). Finally, for intensive longitudinal data with a multilevel structure, Schuurman and Hamaker (2019) introduced the measurement error vector autoregressive model, to parse out within- and between-person reliabilities.

Another proposed solution for obviating problems associated with pursuing high Cronbach's alpha in scale development is evaluating measures for their concurrent and predictive criterion validity. In psychology, concurrent validity is understood as how well a measure correlates with another measure at the same point of measurement, whereas predictive validity reflects how well a measure predicts a future behavior or state (i.e., at a later measurement occasion). Concurrent and predictive validity offer a common ground to evaluate both single- and multiple-item measures. Because unreliable measures cannot yield adequate validity, an equally valid single-item measure can be regarded as sufficiently reliable as a multiple-item measure (Gorsuch & McFarland, 1972). In fact, Cronbach (1960) stated that we should not be discouraged from using a measure with low reliability if it has strong predictive validity.

Few studies to date, however, have directly compared concurrent and predictive validity of a single-item measure with a multiple-item scale. In marketing research, Bergkvist and Rossiter (2007) demonstrated that single-item marketing measures of attitudes toward the ad and the brand exhibited concurrent validity that was equivalent to their multiple-item counterparts and posited that the equivalence would hold for other free-standing, tailor-made single-item measures. Bergkvist (2015) also later showed that the same single items and their multiple-item counterparts held predictive validity with minimal statistical differences. In sports management research, Kwon and Trail (2005) compared the predictive validity of single- and multiple-item measures for affective commitment to a team and team identification. Their findings were mixed; while the affective commitment to a team scale explained more variance in behavioral items than its single-item counterpart, the opposite was true for team identification. Similarly, the use of single items to assess self-esteem (Robins et al., 2001) and social support (Slavin et al., 2020) was supported in general adult populations but not in a sample of children and pregnant women, respectively. Such equivocal findings suggest potential context specificity for the predictive validity of both single- and multiple-item measures. A particular single-item measure may perform well in one setting, whereas the same single item may yield poor results in another (Diamantopoulos et al., 2012).

Nonetheless, evidence for concurrent and predictive validity of single-item measures in several areas of psychological research has been growing. Eddy et al. (2019) examined concurrent and predictive validity of single-item measures for teacher stress and coping. After controlling for covariates, they found that stress and coping predicted concurrent and future emotional exhaustion and demonstrated that the single coping item was sensitive enough to detect intervention effects. More broadly, another example of the use of a well-validated single-item measure is self-rated health in medicine, which has shown to predict mortality across numerous studies (e.g., Finch et al., 2002; Ganna & Ingelsson, 2015; Mossey & Shapiro, 1982). Similarly, a scoping review of a single item measuring self-rated mental health has supported its use in epidemiological studies (Ahmad et al., 2014), and support for the use of single items has been found for mood (Russell et al., 1989; van Rijsbergen et al., 2012), narcissism (Konrath et al., 2014), and personality (Konstabel et al., 2017; Spörrle & Bekk, 2014; Woods & Hampson, 2005). Taken together, these findings provide support for the use of well-designed single-item measures in research and practice.

In addition, there are practical advantages to using single-item measures. Time is a precious commodity for both researchers and respondents, and replacing a multiple-item scale with a single-item measure can save time for both parties. Multiple-item measures also have relatively high participant burden as respondents need to devote more cognitive effort to answering them. As previously noted, such survey fatigue can lower overall response quality and create unnecessary measurement error (Viswanathan & Kayande, 2012). Time and burden constraints are especially true for ecological momentary assessments (EMAs), where participants are often asked to repeatedly answer the same survey across multiple time points. In such a format, including full multiple-item scales to measure specific constructs is especially impractical. For instance, employing the Beck Depression Inventory (Beck et al., 1996) to assess depression would occupy 21 items of a potential EMA survey. Given that these surveys often attempt to achieve coverage across multiple clinical constructs and behaviors, such an approach would likely prove unwieldy. Instead, if constructs of interest are clearly defined and can be captured with fewer items, researchers can choose to either capture additional constructs using the same number of items or lower the number of items to reduce participant burden. In short, a well-validated single-item measure might be able to provide equally accurate and relevant information in a shorter amount of time than its multiple-item counterpart.

The goal of the present study was to interrogate the assumption that single-item measures are not reliable enough to exhibit adequate validity. Concurrent and predictive validity offer the common ground to compare the psychometric properties of single- and multiple-item measures. In the present study, we therefore evaluated the concurrent and predictive validity of single- and multiple-item measures using EMA data. We hypothesized that single-item measures in EMA would exhibit adequate validity comparable to their multiple-item counterparts.

Method

Participants

The present study represents a secondary analysis of previously published data (Fisher et al., 2019), and no separate power analyses were conducted for the present study. Participants ($N = 45$) were individuals with primary diagnoses of generalized anxiety disorder (GAD, $n = 23$), major depressive disorder (MDD, $n = 11$), or both ($n = 11$) who were deemed eligible for an open trial of a personalized cognitive-behavioral intervention for mood and anxiety disorders. Participants were predominantly female ($n = 30$, 65%) and White ($n = 21$, 46%).

Measures

Experience Sampling Survey. For each survey, participants rated their experience of each item over the preceding hours using a 0 to 100 visual analog slider with the anchors *not at all* and *as much as possible* for the 0 and 100 positions, respectively. The sliders were positioned at 50 by default, and participants were required to move the slider to provide their ratings. Surveys contained the extant symptoms of the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5*; American Psychiatric Association, 2013) criteria for GAD and MDD (*down and depressed*, *hopeless*, *loss of interest or pleasure*, *worthless or guilty*, *worried*, *restless*, *irritable*, *difficulty concentrating*, *muscle tension*, *fatigued*, and *anhedonia*) and 10 additional single items: *positive*, *energetic*, *enthusiastic*, *content*, *angry*, *afraid*, *dwelled on the past, avoided people*, *avoided activities*, and *sought reassurance*.

Procedure

In the parent study, we invited individuals with symptomatic experiences consistent with GAD and MDD to contact the last author's laboratory at the University of California, Berkeley. After completing a structured clinical interview to establish diagnosis, eligible participants provided intensive repeated measures data via EMA. The survey system prompted participants through text messages to complete survey questions 4 times per day during their waking hours for 30 days. Participants received surveys about every 4 hours and the exact time of the ping was randomized within a 30-minute window. The average completion rate was 80% ($SD = 8\%$; minimum: 63%, maximum: 95%), and participants provided 5,020 observations in total. More detailed procedures from the parent study are outlined in Fisher et al. (2019).

Analytic Plan

All analyses were conducted in R (Version 4.0.3; R Core Team, 2020).

Factor Analysis. As the present study was a secondary analysis, we were constrained by decisions made in the parent study's design (Fisher et al., 2019). We did not have full multiple-item scales in our EMA survey and instead created three face-valid, multiple-item measures of our own: positive affect (*positive*, *enthusiastic*, *energetic*, and *content*), negative affect (*down and depressed*, *afraid*, *angry*, and *worried*), and depression (*down and depressed*, *anhedonia*, *hopeless*, and *guilty*). Before constructing prediction models with our multiple-item measures, we used the lavaan package (Rosseel, 2012) to test their putative unidimensionality (i.e., single-factor structure). Because our EMA data set had ≥ 90 time points per participant, we used multilevel confirmatory factor analysis (MCFA) to account for the nested structure (Huang, 2018). We examined confirmatory fit indices (CFIs; Bentler & Bonett, 1980) and standardized root mean square residuals (SRMRs) to evaluate the goodness of fit of factor structures. We used the robust maximum likelihood (MLR) estimator to appropriately adjust standard errors and chi-square test statistics for the non-normality of our observations. We also computed a multilevel Cronbach's alpha to assess scale reliability for each multiple-item measure. Furthermore, from the same multilevel factor models, we extracted conservative estimates of single-item reliability following the procedure outlined by Weiss (1976).

Repeated Measures Correlations. We used the rmcorr package (Bakdash & Marusich, 2017) to evaluate the concurrent validity of single- and multiple-item measures. Given our intensive longitudinal data collection, repeated measures correlation was appropriate for determining common within-person association for paired measures assessed on multiple occasions for multiple individuals. We computed repeated measures correlations for all our single- and multiple-item measures except between multiple-item measures and their component single items and between multiple-item measures with shared component single items.

Prediction Models. To compare the predictive validity of single- and multiple-item measures, we used the lme4 package (Bates et al., 2015) to construct multilevel cross-lagged models, where the outcome variable from time point T2 was regressed on the cross-lagged single- or multiple-item predictor, the autocorrelated outcome variable from time point T1, and linear time. Outcome variables were selected for their potential clinical utility (depression, positive affect, and negative affect) and observability (*avoided activities*, *avoided people*, and *dwelled on the past*) and included a combination of single- and multiple-item measures. In all our multilevel models, we included a random intercept and a random slope of linear time.

We made a total of eight sets of comparisons: positive affect → depression, depression → positive affect, negative affect → *avoided activities*, depression → *avoided activities*, negative affect → *avoided people*, depression → *avoided people*, negative affect → *dwelled on the past*, and depression → *dwelled on the past*. These tests were selected for their putative clinical utility and face validity. In each set of comparison, we examined the effect sizes of a multiple-item predictor and four single-item predictors. For example, in the positive affect → depression set of comparisons, we compared the effect size of the positive affect scale with those of the *positive*, *enthusiastic*, *energetic*, and *content* single items. A predictor with a larger effect size was determined to exhibit greater predictive validity.

Results

Unidimensionality and Reliability of Single- and Multiple-Item Measures

Single-factor structures for all three multiple-item measures exhibited acceptable to excellent fit indices: positive affect, $\chi^2(4, N = 5{,}020) = 116.70, p < .001$, CFI = .96, $\text{SRMR}_{\text{within}} = .038$, $\text{SRMR}_{\text{between}} = .089$; negative affect, $\chi^2(4, N = 5{,}020) = 32.01, p < .001$, CFI = .97, $\text{SRMR}_{\text{within}} = .017$, $\text{SRMR}_{\text{within}} = .024$; and depression, $\chi^2(4, N = 5{,}020) = 34.93, p < .001$, CFI = .99, $\text{SRMR}_{\text{within}} = .016$, $\text{SRMR}_{\text{within}} = .039$. Multilevel coefficient alphas also ranged from acceptable to excellent: positive affect ($\alpha_{\text{within}} = .82$, $\alpha_{\text{between}} = .90$), negative affect ($\alpha_{\text{within}} = .71$, $\alpha_{\text{between}} = .90$), and depression ($\alpha_{\text{within}} = .80$, $\alpha_{\text{between}} = .94$).

The communalities of the single items from the positive affect factor model were: *positive* ($\alpha_{\text{within}} = .60$, $\alpha_{\text{between}} = .95$), *enthusiastic* ($\alpha_{\text{within}} = .60$, $\alpha_{\text{between}} = .70$), *energetic* ($\alpha_{\text{within}} = .42$, $\alpha_{\text{between}} = .41$), and *content* ($\alpha_{\text{within}} = .50$, $\alpha_{\text{between}} = .74$). For the negative affect factor, they were: *down and depressed* ($\alpha_{\text{within}} = .45$, $\alpha_{\text{between}} = .53$), *afraid* ($\alpha_{\text{within}} = .37$, $\alpha_{\text{between}} = .69$), *angry* ($\alpha_{\text{within}} = .26$, $\alpha_{\text{between}} = .78$), and *worried* ($\alpha_{\text{within}} = .45$, $\alpha_{\text{between}} = .74$). Finally, for the depression factor, they were *down and depressed* ($\alpha_{\text{within}} = .62$, $\alpha_{\text{between}} = .85$), *anhedonia* ($\alpha_{\text{within}} = .39$, $\alpha_{\text{between}} = .69$), *hopeless* ($\alpha_{\text{within}} = .54$, $\alpha_{\text{between}} = .77$), and *guilty* ($\alpha_{\text{within}} = .47$, $\alpha_{\text{between}} = .86$). The reliabilities of single items ranged from 37% to 78% of their respective multiple-item counterparts at the within-person level and from 46% to 106% at the between-person level.

Concurrent Validity of Single Items

Repeated measures correlations for the single- and multiple-item measures are shown in Table 1. Both the positive affect scale and its four-component single items (*positive*, *enthusiastic*, *energetic*, and *content*) were significantly correlated with the negative affect (scale: $r = -.45$; single items: rs = −.43, −.33, −.24, −.44) and depression (scale: $r = -.53$; single items: rs = −.48, −.40, −.34, −.49) scales. Similarly, the negative affect scale and its four-component single items (*down and depressed*, *afraid*, *angry*, and *worried*) were significantly correlated with the positive affect scale (scale: $r = -.45$; single items: rs = −49, −.24, −.22, −.34) and the depression scale (scale: $r = .78$; single items excluding *down and depressed:* rs = .48, .42, .51). Finally, the depression scale and its four-component single items (*down and depressed*, *anhedonia*, *hopeless*, and *guilty*) were significantly correlated with the positive affect scale (scale: $r = -.53$; single items: rs = −.49, −.46, −.39, −.34) and negative affect scale (scale: $r = .78$; single items excluding *down and depressed:* rs = .49, .61, .59).

Compared with the positive affect scale, relative sizes of correlations of *positive*, *enthusiastic*, *energetic*, and *content* were 95%, 73%, 53%, and 98% of the negative affect scale and 91%, 76%, 64%, and 92% of the depression scale. Compared with the negative affect scale, relative sizes of correlations of *down and depressed*, *afraid*, *angry*, and *worried* were 109%, 53%, 49%, and 76% of the positive affect scale and 61%, 54%, and 65% of the depression scale (excluding *down and depressed*). Compared with the depression scale, relative sizes of correlations of *down and*

Table 1. Repeated Measures Correlations for Single- and Multiple-Item Measures.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Positive affect	—												
2. Negative affect	−.45	—											
3. Depression	−.53	.78	—										
4. Positive	—	−.43	−.48	—									
5. Enthusiastic	—	−.33	−.40	.58	—								
6. Energetic	—	−.24	−.34	.47	.58	—							
7. Content	—	−.44	−.49	.61	.52	.41	—						
8. Angry	−.22	—	.42	−.22	−.17	−.08	−.26	—					
9. Afraid	−.24	—	.48	−.24	−.16	−.11	−.25	.30	—				
10. Down and depressed	−.49	—	—	−.44	−.36	−.32	−.44	.38	.39	—			
11. Worried	−.34	—	.51	−.33	−.26	−.17	−.34	.32	.43	.44	—		
12. Anhedonia	−.46	.49	—	−.40	−.36	−.32	−.40	.27	.29	.52	.33	—	
13. Hopeless	−.39	.61	—	−.35	−.29	−.23	−.37	.36	.42	.57	.42	.44	—
14. Guilty	−.34	.59	—	−.32	−.24	−.20	−.34	.33	.43	.53	.42	.40	.53

Note. All *p*s $< .001$.

depressed, *anhedonia*, *hopeless*, and *guilty* were 92%, 87%, 74%, and 64% of the positive affect scale and 63%, 78%, and 76% of the negative affect scale (excluding *down and depressed*). In summary, compared with their multiple-item counterparts, relative sizes of correlations of single items with the scales (M = 75%, SD = 17%) ranged from 49% (*angry* and positive affect) to 109% (*down and depressed* and positive affect).

Predictive Validity of Single Items

Fixed effects for all the models are shown in Tables 2 to 5, and each table contains two sets of single- and multiple-item model comparisons. One of the comparisons in Table 2, for example, is between the depression scale and its four-component single-item models. The cross-lagged depression scale at T1 (d = −0.22) predicted the positive affect scale at T2, as did each of the four cross-lagged single items at T1, *down and depressed* (d = −0.20; 91% of the effect size for the predictive effect of the depression scale), *anhedonia* (d = −0.16; 73%), *hopeless* (d = −0.13; 59%), *and guilty* (d = −0.10; 45%). In all five models, the autocorrelated positive affect scale at T1 predicted the positive affect scale at T2. The effect size of the depression scale predictor was larger than those of the four single items.

To summarize, in 27 of 29 (93%) unique single-item predictor models, single items demonstrated significant predictive validity. In one of our eight (13%) sets of single- and multiple-item predictor comparisons, there was a single-item predictor with a larger effect size than its multiple-item counterpart. *Down and depressed* (d = 0.18) was a better predictor of *avoided people* than the negative affect scale. Compared with their multiple-item counterparts, effect sizes for the predictive effect of single items ranged from 27% (*afraid* → *avoided people*) to 120% (*down and depressed* → *avoided people*) of the negative affect scale (M = 70%, SD = 21%).

Discussion

The present study offers evidence supporting the concurrent and predictive validity of single items in EMA surveys. During a 30-day pretherapy assessment period, we collected intensive repeated measures of mood and anxiety symptoms and affective states 4 times per day. We created three multiple-item measures of positive affect, negative affect, and depression and confirmed their unidimensionality and scale reliability via multilevel confirmatory factor analysis. We then examined repeated measures correlations for single- and multiple-item measures and compared how well single-item measures and their multiple-item counterparts could predict future behavior or states. Our results demonstrated that single-item measures can exhibit adequate concurrent and predictive validity, comparable to multiple-item measures.

Repeated measures correlations indicated that there was a wide range of concurrent validity among the single-item measures used as components for the same multiple-item scale. *Down and depressed* and the depression scale, for example, were comparable in their negative correlations with the positive affect scale, whereas the correlation between *guilty* and the positive affect scale was about 64% of that between the depression and positive affect scales. Similarly, while *positive* and the positive affect scales were comparable in their negative correlations with the negative affect and depression scales, *energetic* was not as well correlated. One possible explanation is that participants had a good understanding of feeling depressed or positive and

Table 2. Fixed Effects of Positive Affect and Depression Scale and Single Items on Each Other.

CL variable	*B*	*SE*	*t*	*p*	Cohen's *d*	BIC
Positive affect → depression						
Positive affect scale	−0.11	0.02	−6.07	<.001	−0.23	26,341
Positive	−0.09	0.01	−5.89	<.001	−0.21	26,342
Enthusiastic	−0.06	0.01	−4.19	<.001	−0.15	26,360
Energetic	−0.03	0.01	−2.42	.015	−0.09	26,371
Content	−0.09	0.02	−5.88	<.001	−0.21	26,343
Depression → positive affect						
Depression scale	−0.08	0.02	−5.03	<.001	−0.22	26,113
Down and depressed	−0.07	0.01	−5.25	<.001	−0.20	26,110
Anhedonia	−0.05	0.01	−4.34	<.001	−0.16	26,119
Hopeless	−0.04	0.01	−3.05	.002	−0.13	26,128
Guilty	−0.03	0.01	−2.33	.020	−0.10	26,132

Note. Each cross-lagged variable represents a separate model. In every model, the autocorrelated outcome variable at T1 significantly predicted the outcome variable at T2 (all $ps < .001$). CL = cross-lagged; BIC = Bayesian information criterion.

Table 3. Fixed Effects of Negative Affect and Depression Scale and Single Items on Avoiding Activities.

CL variable	*B*	*SE*	*t*	*p*	Cohen's *d*	BIC
Negative affect → avoiding activities						
Negative affect scale	0.16	0.03	5.71	<.001	0.23	29,259
Down and depressed	0.12	0.02	6.06	<.001	0.22	29,254
Afraid	0.04	0.02	1.74	.082	0.07	29,288
Angry	0.07	0.02	3.70	<.001	0.14	29,277
Worried	0.09	0.02	4.62	<.001	0.17	29,270
Depression → avoiding activities						
Depression scale	0.20	0.03	7.77	<.001	0.29	29,232
Down and depressed	0.12	0.02	6.06	<.001	0.22	29,254
Anhedonia	0.13	0.02	6.76	<.001	0.24	29,246
Hopeless	0.11	0.02	5.41	<.001	0.20	29,262
Guilty	0.11	0.02	5.35	<.001	0.20	29,263

Note. Each cross-lagged variable represents a separate model. In every model, the autocorrelated outcome variable at T1 significantly predicted the outcome variable at T2 (all $ps < .001$). CL = cross-lagged; BIC = Bayesian information criterion.

were able to holistically report these states using single items without having to consider all different facets of depression and positive affect in the way that they are commonly discussed in the clinical psychology literature. Because the parent study design did not include a single item on feeling negative, we could not conduct a similar comparison with the negative affect scale. In future research, it might be worth investigating whether participants can accurately report other multifaceted constructs like negative affect using single items.

All but two single-item predictor models demonstrated significant predictive validity. Cronbach's alpha, a common measure of internal consistency, cannot be estimated for single items. However, if our single items were so unreliable and prone to random errors, they would not have been able to form such consistent relationships with other constructs at future time points (Gorsuch & McFarland, 1972). Single-item measures might also exhibit especially sound reliability and validity in repeated assessments such as EMAs. Nomothetic studies that aggregate singular responses rely on the participants to have formed a consensus in their understanding of the items. In idiographic approaches, on the contrary, each respondent develops their own, internally consistent understanding of the items as they repeatedly respond to survey prompts. Given that single-item measures can exhibit sound psychometric properties in the right context, we believe that researchers should carefully evaluate their potential utility across study designs and research contexts.

The *afraid* single item was not predictive of future avoidance of people and activities. This result signals that contextual information might be an important consideration for assessing fear-related avoidance. Although some might report feeling afraid of interacting with other people, others

Table 4. Fixed Effects of Negative Affect and Depression Scale and Single Items on Avoiding People.

CL variable	*B*	*SE*	*t*	*p*	Cohen's *d*	BIC
Negative affect → avoiding people						
Negative affect scale	0.11	0.03	3.96	<.001	0.15	29,083
Down and depressed	0.10	0.02	5.06	<.001	0.18	29,073
Afraid	0.02	0.02	1.13	.259	0.04	29,097
Angry	0.04	0.02	2.26	.024	0.08	29,093
Worried	0.05	0.02	2.45	.013	0.09	29,092
Depression → avoiding people						
Depression scale	0.15	0.03	5.85	<.001	0.21	29,065
Down and depressed	0.10	0.02	5.06	<.001	0.18	29,073
Anhedonia	0.07	0.02	4.04	<.001	0.14	29,082
Hopeless	0.07	0.02	3.83	<.001	0.14	29,084
Guilty	0.10	0.02	4.76	<.001	0.17	29,076

Note. Each cross-lagged variable represents a separate model. In every model, the autocorrelated outcome variable at T1 significantly predicted the outcome variable at T2 (all *p*s < .001). CL = cross-lagged; BIC = Bayesian information criterion.

Table 5. Fixed Effects of Negative Affect and Depression Scale and Single Items on Rumination.

CL variable	*B*	*SE*	*t*	*p*	Cohen's *d*	BIC
Negative affect → rumination						
Negative affect scale	0.20	0.02	8.12	<.001	0.29	28,290
Down and depressed	0.13	0.02	7.96	<.001	0.28	28,292
Afraid	0.08	0.02	4.12	<.001	0.15	28,338
Angry	0.06	0.02	3.42	<.001	0.12	28,343
Worried	0.12	0.01	7.00	<.001	0.25	28,306
Depression → rumination						
Depression scale	0.19	0.02	8.46	<.001	0.30	28,284
Down and depressed	0.13	0.02	7.96	<.001	0.28	28,292
Anhedonia	0.08	0.02	5.05	<.001	0.18	28,329
Hopeless	0.12	0.02	7.01	<.001	0.25	28,306
Guilty	0.10	0.02	5.70	<.001	0.20	28,322

Note. Each cross-lagged variable represents a separate model. In every model, the autocorrelated outcome variable at T1 significantly predicted the outcome variable at T2 (all *p*s < .001). CL = cross-lagged; BIC = Bayesian information criterion.

might report feeling afraid of engaging in certain activities. Our survey unfortunately did not distinguish between potentially varying sources of fear, which may have contributed to these null results. A researcher who wants to predict future avoidant behaviors using the fear item thus might get higher prediction accuracy if they specified the target of the fear emotion. This might be especially true for EMA studies as new contextual information gets incorporated into the participants' consideration at different survey timepoints.

Even in comparisons where the multiple-item models outperformed all four-component items, there was at least one single-item predictor with a minimal difference in effect size from the multiple-item predictor. For example, although the positive affect scale predicted depression better than all four-component single items (*positive*, *enthusiastic*, *energetic*, and *content*), the respective differences in Cohen's *d* between the scale and *positive* and *content* were both 0.02. If these minimal differences are not of concern to research outcomes, the researcher might choose to lower relatively high participant burden in EMA studies by forgoing the positive affect scale in favor of either of the two single items with comparable predictive validity. One notable exception was the depression and avoiding activities model. In this set of comparisons, the multiple-item depression score consistently outperformed the single-item components (*down and depressed*, *anhedonia*, *hopeless*, and *guilty*) with relatively sizable differences in Cohen's *d* (0.07, 0.05, 0.09, and 0.09, respectively). This might point to avoidance of activities likely stemming from the combination of depressive symptoms rather than one specific symptom.

Although the depression scale and *down and depressed* predicted future positive affect, avoidance, and rumination comparably, there was a notable difference in effect size for avoidance of future activities. These results support the idea of context specificity of predictive validity previously noted by Diamantopoulos et al. (2012). If the research objectives include forecasting positive affect, avoidance of people, and rumination, the *down and depressed* single item will not fall too far behind the full depression scale in the present data. However, researchers should not assume one measure will always perform comparably with another as our results demonstrated the superiority of the depression scale in predicting avoidance of activities. In the same vein, single items that performed relatively poorly in the present study might also exhibit greater predictive validity in answering other research questions. Furthermore, the wording of single items designed to assess the same construct might also matter. When assessing sad mood in EMA surveys, for example, a researcher might ask participants to rate how *sad*, *gloomy*, or *down and depressed* they feel. It remains an empirical question which wording for sad mood and other target constructs would provide the most predictive utility for researchers. Thus, if the prediction accuracy is of concern, researchers should carefully consider the relationship between the predictor and criterion constructs before deciding on the prediction models. Nonetheless, predictive validity demonstrated by the *down and depressed* single item and others underscores the likelihood that carefully designed single items in intensive longitudinal data can exhibit adequate validity while reducing research participation burden.

The discussion of psychometric properties of symptom-specific single items might be timelier than ever in clinical psychology. The field has been moving toward acknowledging substantial heterogeneity in mental health diagnoses both between and within subjects (Fisher et al., 2018, 2019; Wright & Simms, 2016). That is, two people rarely experience identical mental health symptoms even for the same diagnosis, and even within a person, their symptoms fluctuate overtime (Fisher & Bosley, 2020; Howe et al., 2020). The possible fallibility of relying on multiple-item measures to capture overall symptom severity is in line with Rossiter's (2002) terminology. If a target construct is multifaceted like most, if not all, mental health diagnoses, it cannot be classified as a concrete singular construct by definition. Instead, more granular approaches to understanding psychopathology, potentially involving singular and concrete, symptom-specific single item, might pave the way for innovative ways to understand and treat a complex network of symptoms.

The quality of single item thus might vary widely by the appropriateness of its target construct, namely, whether it is concrete and singular enough for the participants to rate without confusion. For example, a single item assessing sleep problems using EMA surveys might not be specific enough. A participant might report that they experienced sleep problems one night because of sleep fragmentation (i.e., they could not stay asleep), whereas they might report the same degree of sleep problems the next night because of long sleep latency (i.e., they could not fall asleep). Instead, separate single items for these constructs would appropriately differentiate the two. Like in any other survey methods, some construct, like a clinical diagnosis, might also be never appropriate to measure with a single item in EMA surveys because the construct is too broad for both researchers and respondents.

Limitations

Although the present study directly compares criterion validity of single- and multiple-item measures in EMA surveys, a few limitations need to be noted. First, the present study was limited by the parent study design to compare single- and multiple-item measures of the same constructs. Instead, we demonstrate that emotion-specific single items of lower order can exhibit predictive validity comparable with higher-order multiple-item measures in certain contexts. The *down and depressed* single item, for example, constitutes only one facet of the higher-order depression scale and the two do not necessarily measure the same construct. To address this limitation, a future study could ask participants to rate a higher-order construct using a single item and compare its criterion validity with that of the well-validated scale collected in the same EMA response. For instance, participants could respond to a single negative affect item as well as the well-validated negative affect scale from the Positive and Negative Affect Schedule (Mackinnon et al., 1999). Such original data set would allow for the comparison of criterion validity of single- and multiple-item measures of the same construct.

Second, while we highlighted the importance of criterion validity in evaluating the psychometric properties of single- and multiple-item measures, it might be worth noting that high validity of multiple-item measures does not always equate with high internal consistency among the items in the scale. That is, criterion validity does not indicate that the scale is concrete and singular. In fact, the bad items that are predictors of the criterion on their own can increase the predictive validity of the scale if their variation is correlated with the variation in the criterion. Future studies thus should not solely evaluate multiple-item measures for their criterion validity but also check for their unidimensionality and internal consistency as we did for the multiple-item measures in the present study.

Third, while we proposed to examine criterion validity in the absence of Cronbach's alpha for single items, there are other metrics of reliability that could be estimated for single items (e.g., Nunnally & Bernstein, 1994; Schuurman & Hamaker, 2019; Weiss, 1976). Furthermore, in intensive longitudinal data, the definition of reliability can be expanded beyond the within-person consistency to include the precision and stability of person-specific statistics, such as mean, standard deviation, autoregression, and correlation (Wright & Zimmerman, 2019). Future studies comparing the consequences of using single- and multiple-item measures on the reliability of such person-specific measures could further inform the utility of single items.

Conclusion

The present study provides novel evidence supporting concurrent and predictive validity of single-item measures in EMA surveys and therefore their use in research and practice. Criterion validity is an especially important metric for evaluating psychometric properties because it offers a common testing ground for both single- and multiple-item measures. Concurrently and predictively valid single items will undoubtedly prove useful in EMA studies where full scales may cause unnecessary participant burden and lower their response quality.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Jiyoung Song https://orcid.org/0000-0002-7157-8198

Joshua R. Oltmanns https://orcid.org/0000-0001-6670-6995

Aaron J. Fisher https://orcid.org/0000-0001-9754-4618

References

Ahmad, F., Jhajj, A. K., Stewart, D. E., Burghardt, M., & Bierman, A. S. (2014). Single item measures of self-rated mental health: A scoping review. *BMC Health Services Research*, *14*(1), 1–11. https://doi.org/10.1186/1472-6963-14-398

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, *8*, 456. https://doi.org/10.3389/fpsyg.2017.00456

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, A. T., Steer, R. A., & Brown, G. (1996). *Beck Depression Inventory–II*. The Psychological Corporation.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bergkvist, L. (2015). Appropriate use of single-item measures is here to stay. *Marketing Letters*, *26*(3), 245–255. https://doi.org/10.1007/s11002-014-9325-y

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same construct. *Journal of Marketing Research*, *44*(2), 175–184. https://doi.org/10.1509/jmkr.44.2.175

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*(1), 64–73. https://doi.org/10.2307/3150876

Clark, L. A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 187–203). American Psychological Association. https://doi.org/10.1037/14805-012

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed., Vol. 26). SAGE.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, *40*(3), 434–449. https://doi.org/10.1007/s11747-011-0300-3

Eddy, C. L., Herman, K. C., & Reinke, W. M. (2019). Single-item teacher stress and coping measures: Concurrent and predictive validity and sensitivity to change. *Journal of School Psychology*, *76*, 17–32. https://doi.org/10.1016/j.jsp.2019.05.001

Finch, B. K., Hummer, R. A., Reindl, M., & Vega, W. A. (2002). Validity of self-rated health among Latino(a)s. *American Journal of Epidemiology*, *155*(8), 755–759. https://doi.org/10.1093/aje/155.8.755

Fisher, A. J., & Bosley, H. G. (2020). Identifying the presence and timing of discrete mood states prior to therapy. *Behaviour Research and Therapy*, *128*, 103596. https://doi.org/10.1016/j.brat.2020.103596

Fisher, A. J., Bosley, H. G., Fernandez, K. C., Reeves, J. W., Soyster, P. D., Diamond, A. E., & Barkin, J. (2019). Open trial of a personalized modular treatment for mood and anxiety. *Behaviour Research and Therapy*, *116*, 69–79. https://doi.org/10.1016/j.brat.2019.01.010

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(72), 1–11. https://doi.org/10.1016/10.1186/s12916-015-0325-4

Ganna, A., & Ingelsson, E. (2015). 5 year mortality predictors in 498 103 UK Biobank participants: A prospective population-based study. *The Lancet*, *386*(9993), 533–540. https://doi.org/10.1016/S0140-6736(15)60175-1

Gorsuch, R. L., & McFarland, S. G. (1972). Single vs. multiple-item scales for measuring religious values. *Journal for the Scientific Study of Religion*, *11*(1), 53–64. https://doi.org/10.2307/1384298

Howe, E., Bosley, H. G., & Fisher, A. J. (2020). Idiographic network analysis of discrete mood states prior to treatment. *Counselling and Psychotherapy Research*, *20*(3), 470–478. https://doi.org/10.1002/capr.12295

Huang, F. L. (2018). *Conducting multilevel confirmatory factor analysis using R* [Unpublished manuscript]. Department of Psychology, University of Missouri.

Jacoby, J. (1978). Consumer research: A state of the art review. *Journal of Marketing*, *42*(2), 87–96. https://doi.org/10.2307/1249890

Konrath, S., Meier, B. P., & Bushman, B. J. (2014). Development and validation of the single item narcissism scale (SINS). *PLOS ONE*, *9*(8), Article e103469. https://doi.org/10.1371/journal.pone.0103469

Konstabel, K., Lönnqvist, J. E., Leikas, S., García Velázquez, R., Qin, H., Verkasalo, M., & Walkowitz, G. (2017). Measuring single constructs by single items: Constructing an even shorter version of the "Short Five" personality inventory. *PLOS ONE*, *12*(8), Article e0182714. https://doi.org/10.1371/journal.pone.0182714

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Kwon, H., & Trail, G. (2005). The feasibility of single-item measures in sport loyalty research. *Sport Management Review*, *8*(1), 69–88. https://doi.org/10.1016/S1441-3523(05)70033-4

Mackinnon, A., Jorm, A. F., Christensen, H., Korten, A. E., Jacomb, P. A., & Rodgers, B. (1999). A short form of the Positive and Negative Affect Schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample. *Personality and Individual Differences*, *27*(3), 405–416. https://doi.org/10.1016/S0191-8869(98)00251-7

Mossey, J. M., & Shapiro, E. (1982). Self-rated health: A predictor of mortality among the elderly. *American Journal of Public Health*, *72*(8), 800–808. https://doi.org/10.2105/ajph.72.8.800

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill Education.

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, *16*(1), 6–17. https://doi.org/10.2307/3150868

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*(2), 151–161. https://doi.org/10.1177/0146167201272002

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*(4), 305–335. https://doi.org/10.1016/S0167-8116(02)00097-6

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502. https://doi.org/10.1037/0022-3514.57.3.493

Sackett, P. R., & Larson, J. R., Jr. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 419–489). Consulting Psychologists Press.

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

Slavin, V., Creedy, D. K., & Gamble, J. (2020). Single item measure of social supports: Evaluation of construct validity during pregnancy. *Journal of Affective Disorders*, *272*, 91–97. https://doi.org/10.1016/j.jad.2020.03.109

Spörrle, M., & Bekk, M. (2014). Meta-analytic guidelines for evaluating single-item reliabilities of personality instruments. *Assessment*, *21*(3), 272–285. https://doi.org/10.1177/1073191113498267

van Rijsbergen, G. D., Bockting, C. L., Berking, M., Koeter, M. W., & Schene, A. H. (2012). Can a one-item mood scale do the trick? Predicting relapse over 5.5-years in recurrent depression. *PLOS ONE*, *7*(10), e46796. https://doi.org/10.1371/journal.pone.0046796

Viswanathan, M. (2005). *Measurement error and research design*. SAGE.

Viswanathan, M., & Kayande, U. (2012). Commentary on "Common method bias in marketing: Causes, mechanisms, and procedural remedies". *Journal of Retailing*, *88*(4), 556–562. https://doi.org/10.1016/j.jretai.2012.10.003

Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods*, *4*(4), 361–375. https://doi.org/10.1177/109442810144003

Weiss, D. J. (1976). Multivariate procedures. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 327–362). Rand McNally College.

Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality*, *19*(5), 373–390. https://doi.org/10.1002/per.542

Wright, A. G., & Simms, L. J. (2016). Stability and fluctuation of personality disorder features in daily life. *Journal of Abnormal Psychology*, *125*(5), 641–656. https://doi.org/10.1037/abn0000169

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