Editorial

Single Item Measures in Psychological Science

A Call to Action

Mark S. Allen[1], Dragos Iliescu[2], and Samuel Greiff[3]

[1]School of Psychology, University of Wollongong, NSW, Australia
[2]Faculty of Psychology and Educational Sciences, University of Bucharest, Romania
[3]Department of Behavioural and Cognitive Sciences, University of Luxembourg, Luxembourg

Single-item measures have a bad reputation. For a long time, adopting single-item measures was considered one of the surest methods of receiving a letter of rejection from journal editors (Wanous et al., 1997). As one research team noted, "it is virtually impossible to get a journal article accepted … unless it includes multiple-item measures of the main constructs" (Bergkvist & Rossiter, 2007, p. 175). However, a series of articles published in the late 1990s and 2000s began to challenge the conventional view that single-item measures are an unsound approach to measuring cognitive and affective outcomes (Bergkvist & Rossiter, 2007; Fuchs & Diamantopoulos, 2009; Jordan & Turner, 2008; Loo, 2002; Nagy, 2002; Wanous et al., 1997). These articles did much to alleviate the stigma surrounding single-item measures, but even today, many researchers remain unconvinced that single-item measures can provide valid and reliable assessments of important psychological phenomena.

Of course, there are many instances in which single-item measures would be a poor choice – for example, in research aiming to capture the breadth of human personality or emotion. However, when a construct is unambiguous or narrow in scope, the use of single items can be appropriate and should not necessarily be considered unsound (Wanous et al., 1997). The last few decades have seen a marked increase in the use of large national-level panel data in psychological research. Given the considerable volume of data and the diversity of constructs included in these panel surveys, it is often necessary to measure psychological constructs using just a few or even only one item. For example, the Household, Income and Labour Dynamics in Australia Survey (HILDA; Watson & Wooden, 2021) assesses body weight satisfaction using the single item "How satisfied are you with your current weight?" with response categories of 1 (= *very satisfied*), 2 (= *satisfied*), 3 (= *neither satisfied nor dissatisfied*), 4 (= *dissatisfied*), and 5 (= *very dissatisfied*). Although there are multi-item measures of body satisfaction available, on face value, there is no reason to think that this single item does not adequately capture a person's general satisfaction with their body weight. The increasing use of large panel surveys in psychological research means that now more than ever, it is essential to ensure that single-item measures are valid and reliable.

Arguments For and Against Single-Item Measures

Arguments Against Single-Item Measures

Much previous work has discussed the advantages and disadvantages of single item measures. Arguments offered against the use of single-item measures have often been convoluted and are not necessarily convincing from a theoretical point of view. Two such arguments stand out: the assertion that single-item measures have lower (or uncertain) reliability and the assertion that single-item measures lack the capacity for finer-grained assessment (for instance, by mere range restriction given that only one item can be scored).

The first criticism of single-item measures is that estimation of measurement error will not follow the prescribed model that relies on intercorrelations of a scale's components as an estimation of reliability (i.e., the internal consistency approach). That is, without different components of measurement (i.e., other items), single-item measures cannot be subjected to the statistical procedures that fall under the umbrella of "internal consistency." Therefore, alternative methods, which are often cumbersome and time-consuming but still feasible and established, need to be considered. For instance, test-retest reliability (i.e., score stability) can be computed for theoretically stable

constructs, but this is more challenging as it requires a dedicated design with (at least) two measurement points. As most psychological research continues to be cross-sectional (which is an issue in and of itself), this generates a potential problem for estimating scale reliability in cross-sectional studies that might want to include single-item measures. The argument follows that because single-item measures cannot be compared to corresponding items (that capture the same construct), they are more vulnerable to measurement error (Fuchs & Diamantopoulos, 2009; Oshagbemi, 1999). This is based on the Spearman-Brown prophecy – the statistical effect through which measurement error in the total scale score of a multi-item scale decreases as random measurement errors cancel each other out when averaged across items (while true construct variance incrementally adds up).

The hard-line argument is that reliability of single-item measures is simply *lower*, which makes them unsuitable for use. The softer argument is that reliability of single-item measures is simply *unknown* in most cases. This is a lesser, albeit still valid argument that, in many cases, might contribute to researchers concluding that single-item measures are unsuitable for use. Indeed, for cross-sectional research, reliability estimates for single-item measures cannot be computed, and this might be a problem for some statistical applications (e.g., estimation of standard error of measurement for decisions, disattenuation of correlations). In addition, estimates of score stability are not always possible. For example, test-
retest reliability cannot be computed for cognitive and affective outcomes that are predicted to be variable over time (e.g., emotion, mood).

The second argument against single-item measures is that complex psychological constructs cannot be adequately captured using a single item. This argument relates to content validity and also has two components. The first is that for more sophisticated constructs with multidimensional content or a multitude of behavioural expressions (e.g., a personality trait), one item cannot cover sufficient territory of the target construct to be considered valid when compared to a multi-item measure. This is a fair point, and few would claim that a single item could adequately capture the breadth of human personality or emotion. Therefore, the second argument typically focuses on the lack of response categories on single-item measures. That is, multiple items capture more information and therefore allow for more fine-grained distinctions between individuals (Bergkvist & Rossiter, 2007). In this instance, it is not multiple items that make the scale better, but rather, the greater number of response categories. In other words, the same improvement could be achieved (theoretically speaking) by providing more response categories on the single item (e.g., a 7-point scale in preference to a 5-point scale). However, there is little evidence that adding more response categories offers a superior measure (see e.g., Dawes, 2008).

Arguments in Favour of Single-Item Measures

The arguments in favor of single-item measures are, in essence arguments surrounding utility and efficiency, combined with strong evidence that single items can indeed be valid reflections of the underlying construct of interest. Four specific arguments stand out as important when considering the use of single-item measures.

The most obvious benefit of single-item measures is that they are more parsimonious in terms of administration time. They are therefore more appropriate for use in time-restricted conditions. Of course, time-restricted conditions are abundant both in research and in practice. This is particularly important when it comes to large panel surveys where measures are often administered to hundreds of thousands of participants. Single-item measures are also more suitable for vulnerable populations (e.g., adults with intellectual disabilities or clinical patients) who might not have the cognitive (e.g., attention span) or emotional (e.g., impulse control) resources to sit through longer test-taking sessions. Aside from our preferences as researchers and practitioners, or the time that a test-taker might objectively be able to spend with the administration of a measure, we also have an ethical obligation to not waste the time of individuals who participate in testing sessions with superfluous questioning. That single-item measures are less time-consuming has other beneficial effects – for example, they can increase people's willingness to take the time to complete and return a questionnaire (Wanous et al., 1997) or allow researchers to include a larger number of theoretically relevant constructs in research.

The second argument in favor of single-item measures is that they are more satisfying for test-takers. Of course, completing questionnaires is somewhat of a chore, and therefore a shorter scale will undoubtedly be considered more satisfying. Aside from this, test takers can often find multi-item measures repetitious and responding to similarly worded questions to be tedious and even infuriating. As one test-taker once commented to me (first author) after completing a validated 16-item measure of attribution: "...it was quite annoying, why did you ask the same questions over and over again?". This example illustrates a common problem in scale development: that researchers are developing multi-item measures for constructs that are narrow in scope and where a single-item measure will probably suffice. When using scales that contain little breadth, respondents can resent being asked questions that appear repetitious (Wanous et al., 1997). This frustration could

even affect participant responses, such as causing confusion (e.g., "am I supposed to give different answers to what is essentially the same question?"), or less time and effort in answering items (e.g., "all the questions are basically the same, so I guess I will just score 4 for everything"). It is important to note that asking the same question repeatedly is no better than asking it once.

The third argument in favor of single-item measures is that they can reduce data processing costs (Bergkvist & Rossiter, 2007). Shorter measures mean lower costs in preparing digital forms for data collection and less sophisticated programs for collating these data. This is an exponential benefit in case of those projects where data is not collected through digital/computerized channels, but rather in paper-and-pencil format, and where simple data input (including double-checking) can raise costs significantly and bring with it significant opportunities for imputation errors. The fourth argument in favor of single-item measures is that they can be less ambiguous in their measurement of the construct of interest. That is, multiple items provide an opportunity to cover a broader content (in the sense of comprehensive construct coverage), but unfortunately, they also provide more opportunity for the inclusion of items that are ambiguous or unclear (i.e., a greater risk of low face validity), or items that tap into other (related) constructs (i.e., a greater risk of construct contamination). In other words, a scale consisting of one or two "good" items can outperform a scale with multiple items (Bergkvist & Rossiter, 2007).

The case we want to make here is that single-item measures are not automatically inferior to multi-item measures. Given the advantages associated with single-item measures, they can often be viable alternatives and even be superior in many situations. Single-item measures are acceptable when constructs are unidimensional, clearly defined, and narrow in scope (Fuchs & Diamantopoulos, 2009). At face value, it can often be obvious when constructs are too broad to warrant a single item or sufficiently narrow that only a single item is needed. However, there is a middle ground where the feasibility of a single-item measure is unknown. For example, a single-item measure of anxiety (e.g., "how anxious do you feel right now" – scored from 1 = *not at all anxious* to 7 = *extremely anxious*) might be a valid measure of state anxiety. However, the term "anxious" can be interpreted in different ways. For example, a person might report being anxious to mean they are excited and experiencing a state of readiness for an upcoming competition. To capture a more rounded interpretation of a person's emotional state, multiple items using a variety of terms (e.g., worried, concerned, nervous, frightened, uneasy, apprehensive) might be a better approach to capturing the breadth of the emotion. The key point is that until validation tests are done, the trustworthiness of single-item measures will remain unknown.

Types of Validation Tests for Single-Item Measures

Just as for any psychological measure, convincing evidence is needed from different angles to establish the validity of single-item measures. There are some specific approaches that apply to this type of measurement, and the validation process might look somewhat different to validating multi-item measures. We briefly outline some of these approaches.

Face Validity

Face validity is probably the most underused source of validation. It is quite incredible how many new questionnaires are developed that skip this crucial phase. Face validity refers to the clarity or relevance of a test as it appears to participants (Holden, 2010). There are many instances in which items might be valid in one population but be less appropriate for another. For example, the self-report altruism scale (Rushton et al., 1981) includes many examples of altruistic behavior, including the item "I have helped push a stranger's car out of the snow." This item is likely to be a valid measure of altruism in the Canadian sample in which the questionnaire was developed but would likely have low validity in an Australian or African population where there is little or no snow. Similarly, the Big Five Inventory-2 Short-Form (Soto & John, 2017) uses adjectives such as "blue" and "soft heart" that, while common in North America, might cause confusion outside of this region. Just as for multi-item measures, it is critically important for single-item measures to demonstrate face validity. In particular, researchers should aim to establish five components of face validity including: (1) item relevance (item is meaningful and relevant to participants), (2) ease of response (item is not difficult to answer), (3) item ambiguity (item cannot be interpreted different ways), (4) item is not considered distressing or sensitive, and (5) item is not considered judgmental (Connell et al., 2018).

Criterion Validity

Convergent Validity

The most common method of validating single-item measures is through convergent validity with their multi-item counterpart. For instance, a single-item measure of collective efficacy was found to correlate with average scores on a 20-item measure at $r = .69$, $r = .73$, and $r = .74$ across three studies (Bruton et al., 2016). A single-item measure of life satisfaction was also found to correlate with average scores on a 4-item measure at $r = .64$ (disattenuated $r = .80$) (Cheung & Lucas, 2014), and a single-item measure of academic anxiety was found to correlate with average scores

on a 17-item measure at r = .55 (Gogol et al., 2014). The main issue with convergent validity tests is that there is little agreement or guidance on the values that might reflect acceptable convergence. Until a strong argument can be made for particular values, a useful guide might be to consider values similar to those adopted for test-retest reliability, in which r = .90 is indicative of excellent convergent validity, r = .80 indicates good convergent validity, r = .70 indicates acceptable convergent validity, r = .60 indicates questionable convergent validity, and r < .60 indicates poor convergent validity (Greiff & Allen, 2018).

Predictive Validity

In instances where there is no multi-item counterpart to a single-item measure, it can be useful to establish criterion validity through correlations with a theoretical outcome. For example, if a single-item measure of mathematics anxiety predicted subsequent mathematics performance (or "processing efficiency" as predicted by attentional control theory; Eysenck et al., 2007), then this would be considered evidence for the validity of the single-item measure. Much research has supported the validity of single-item measures through correlations with theoretical outcomes measured either concurrently or subsequently (e.g., Eddy et al., 2019; Jovanović, & Lazić, 2020). One key issue is that the single-item measure should predict the target outcome to a pre-specified level (i.e., a predicted effect size). If the observed effect size is smaller than that predicted, then this would be considered evidence against the validity of the new measure. However, studies tend not to present target effect sizes and often accept a statistically significant correlation (dependent on sample size) as supporting predictive validity irrespective of the actual effect size. To test predictive validity as accurately as possible, researchers should preregister their target effect size, or better yet, conduct their validation work using registered report guidelines (see Chambers, 2013; Greiff & Allen, 2018).

Concurrent Validity

Predictive validity can also be considered in combination with convergent validity. If the new single-item measure can predict a theoretical outcome with a similar effect size to its multi-item counterpart, then this is considered further evidence for the validity of the new measure (Bergkvist & Rossiter, 2007). For example, one study found that a multi-item measure of team identification was a better predictor of game-watching behavior (explaining 12.1% more variance) and licensed clothing wearing (explaining 10.7% more variance) than a single-item measure of team identification (Kwon & Trail, 2005). The authors concluded that the multi-item measure was therefore superior. In another study of 11 meta-analyses combining 189 advertising studies, it was found that single-item measures predicted outcomes (attitudes) with almost identical effect sizes to multi-item measures (Ang & Eisend, 2018). This type of validation testing can be extended to the nomological network of a target construct. For instance, by comparing the empirical relation of the single-item measure to the related constructs with the relation that has usually been obtained in the literature (ideally in meta-analysis).

Test-Retest Reliability

For constructs predicted to be relatively stable over time (e.g., attitudes, beliefs), it is also important to establish the reliability of single-item measures. Test-retest reliability involves repeated measures that typically range from one week apart to three months apart. Moreover, the timeframe should be sufficiently long that exact answers to items are not retained in short-term memory, but not so long that dispositions (e.g., attitudes, beliefs, traits) might change naturally over time and thus invalidate the test-retest (Polit, 2014). Correlations between item scores measured at Time 1 and Time 2 can provide insight into scale reliability. For example, one-month and three-month test-retest correlations were explored for 18 single-item measures in 302 organizational workers, with correlations ranging from .46 to .78 at one month and .35 to .77 at three months (Fisher et al., 2016), providing evidence that some single-item measures were more reliable than others. Establishing test-retest reliability is particularly important for single-item measures since additional items are not available to lessen the potential damage incurred by one inconsistent item.

Conclusion

Given the (rather negative) reputation surrounding single-item measures, it is interesting to note that most research published on single-item measures shows that they are often as valid and reliable as their multi-item counterparts (Ahmad et al., 2014; Ang & Eisend, 2018). Perhaps publication bias has played a partial role in this, with unsuccessful validation attempts of single-item measures less likely to be published. But we suspect that researchers are simply developing single-item scales when there is good theoretical reason to suspect that such measures will provide an adequate assessment of the construct of interest. Of note, at *EJPA*, we are more than happy to publish unsuccessful validation attempts (and research with null results more generally), and we particularly encourage authors to submit registered reports. As editors, we can confidently say that we are not inundated with manuscript submissions validating single-item scales (for good examples of single-item scale validation, see Fisher et al., 2016; Gogol et al., 2014). In fact, we would welcome a discussion that might

build on issues raised in this editorial by providing examples of both successful and unsuccessful attempts at developing valid and reliable single-item measures. Thus, this editorial is a call to action for research validating single-item measures, and in particular, those that are already featured in large panel surveys. To conclude, developing and validating multi-item measures for use in research is of little value if single-item measures are being used in practice. In such cases, the more important validation is of the single-item measure, including how closely it approximates a validated multi-item measure. We hope this editorial can stimulate sufficient interest to warrant a special issue of *EJPA* focused on single item validation.

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. http://www.biomedcentral.com/1472-6963/14/398

Ang, L., & Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research, 58*(2), 218–227. https://doi.org/10.2501/JAR-2017-001

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

Bruton, A. M., Mellalieu, S. D., & Shearer, D. A. (2016). Validation of a single-item stem for collective efficacy measurement in sports teams. *International Journal of Sport and Exercise Psychology, 14*(4), 383–401. https://doi.org/10.1080/1612197X.2015.1054853

Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex. *Cortex, 49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: Results from three large samples. *Quality of Life Research, 23*(10), 2809–2818. https://doi.org/10.1007/s11136-014-0726-4

Connell, J., Carlton, J., Grundy, A., Buck, E. T., Keetharuth, A. D., Ricketts, T., Barkham, M., Robotham, D., Rose, D., & Brazier, J. (2018). The importance of content and face validity in instrument development: Lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). *Quality of Life Research, 27*, 1893–1902. https://doi.org/10.1007/s11136-018-1847-y

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50*(1), 61–104. https://doi.org/10.1177/147078530805000106

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

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*(1), 3–23. https://doi.org/10.1037/a0039139

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft, 69*(2), 195–210.

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & Preckel, F. (2014). "My questionnaire is too long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology, 39*(3), 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

Greiff, S., & Allen, M. S. (2018). *EJPA* introduces registered reports as new submission format. *European Journal of Psychological Assessment, 34*(4), 217–219. https://doi.org/10.1027/1015-5759/a000492

Holden, R. B. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encylopedia of psychology* (4th ed., pp. 637–638). Wiley.

Jordan, J. S., & Turner, B. A. (2008). The feasibility of single-item measures for organisational justice. *Measurement in Physical Education and Exercise Science, 12*, 237–257. https://doi.org/10.1080/10913670802349790

Jovanović, V., & Lazić, M. (2020). Is longer always better? A comparison of the validity of single-item versus multiple-item measures of life satisfaction. *Applied Research in Quality of Life, 15*(3), 675–692. https://doi.org/10.1007/s11482-018-9680-6

Kwon, H., & Trail, G. (2005). The feasibility of single-item measures in sport loyalty research. *Sport Management Review, 8*, 68–89. https://doi.org/10.1016/S1441-3523(05)70033-4

Loo, R. (2002). A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology, 17*, 68–75. https://doi.org/10.1108/02683940210415933

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology, 75*, 77–86. https://doi.org/10.1348/096317902167658

Oshagbemi, T. (1999). Overall job satisfaction: How good are single versus multiple-item measures? *Journal of Managerial Psychology, 14*(5), 388–403.

Polit, D. F. (2014). Getting serious about test-retest reliability: A critique of retest research and some recommendations. *Quality of Life Research, 23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

Rushton, J. P., Chrisjohn, R. D., & Fekken, G. C. (1981). The altruistic personality and the self-report altruism scale. *Personality and Individual Differences, 2*(4), 293–302. https://doi.org/10.1016/0191-8869(81)90084-2

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory – 2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are the single item measures? *Journal of Applied Psychology, 82*, 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Watson, N., & Wooden, M. (2021). The Household, Income and Labour Dynamics in Australia (HILDA) Survey. *Journal of Economics and Statistics, 241*(1), 131–141.

Published online January 27, 2022

Mark Allen
School of Psychology
University of Wollongong
Northfields Avenue
Wollongong, NSW 2522
Australia
markal@uow.edu.au