Heliyon 9 (2023) e15280

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Test-retest reliability of single-item assessments of immune fitness, mood, and quality of life

Joris C. Verster [a,b,*], Kiki EW. Mulder [a], Marjolijn CE. Verheul [a], Evi C. van Oostrom [a], Pauline A. Hendriksen [a,b], Andrew Scholey [c], Johan Garssen [a,d]

[a] Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, 3584CG, Utrecht, the Netherlands
[b] Centre for Human Psychopharmacology, Swinburne University, Melbourne, VIC, 3122, Australia
[c] Nutrition Dietetics and Food, School of Clinical Sciences, Monash University, Clayton, VIC, 3168, Australia
[d] Global Centre of Excellence Immunology, Nutricia Danone Research, 3584CT, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Test-retest reliability
Single-item assessments
Mood
Immune fitness
Quality of life
Bland-altman
Intraclass correlation

ABSTRACT

The use of single-item assessments is increasingly important and popular, as these enable quick real-time assessments in clinical practice or research. In this study we investigated the test-retest reliability of single-item assessments of mood ("stress", "anxiety", "depression", "fatigue", "loneliness", "being active", "optimism", and "happiness"), quality of life, and immune fitness in N = 108 participants. The analysis revealed high test-retest correlations between the single-item assessments (r = 0.67 to 0.90), moderate to excellent intraclass correlations (r = 0.672 to 0.889), and the Bland-Altman analysis revealed agreement between all test-retest assessments, except for depression. Taken together, it can be concluded that the single-item assessments of mood, quality and immune fitness have a good test-retest reliability. This strengthens the rationale for using these single item assessments.

1. Introduction

The use of single-item assessments is increasingly important and popular, as these enable quick real-time assessments in clinical practice or research. Single-item scales have several advantages compared to multiple item scales [1,2]. The outcome of single-item assessments is directly available without counting of item scores or recoding. As a result, the use of single-item scale is associated with a reduction in research costs. Since completion of single-item assessments is less time-consuming, their implementation will shorten surveys and clinical assessments. This minimizes the burden for participants and is therefore likely to result in higher response rates.

It is important that such single-item measures have the same validity and reliability as the traditional multiple-item questionnaires. Previously a series of single-item scales have been developed to assess mood and health correlates [2]. The single-item scales showed to be equally effective as the original multiple item scales to assess the corresponding constructs such as quality of life, anxiety, depression, and stress, and they have been used successfully in a series of studies evaluating health and disease [3–11]. The validity and reliability of single item assessment was demonstrated in a large sample of 2489 participants, in which the outcomes of the traditional multiple-item scales were compared to those of the single items [2]. Bland-Altman analysis of agreement revealed that the outcomes of the single-item assessments did not differ from those of the corresponding multiple-item scales [2]. Studies from other

[*] Corresponding author. Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Universiteitsweg 99, 3584CG, Utrecht, the Netherlands.

E-mail address: j.c.verster@uu.nl (J.C. Verster).

https://doi.org/10.1016/j.heliyon.2023.e15280
Received 19 July 2022; Received in revised form 22 March 2023; Accepted 31 March 2023
Available online 5 April 2023

research groups also found that single item ratings were equally sensitive and reliable as traditional multiple item scales that assessed the same construct. This was for example shown for quality of life [12], depression [13], and fatigue [14]. Our group also further investigated a single-item assessment of alcohol hangover severity [15]. For the single item assessing quality of life an intra-class correlation analysis was conducted for two test occasions that revealed an excellent test-retest reliability of 0.87 [12].

It is important to establish that when a scale is completed more than once by the same individual, under the same circumstances, its outcome is consistent/reproducible (and thus reliable). Examining this so-called test-retest reliability is the purpose of the current investigation. To this extent, the single-item scales were completed twice by the same participants. It is essential that the conditions under which the assessments are performed are identical for the specific individual. This is important as the outcomes for some variables under investigation may vary from day to day. The test-retest interval for assessments that are not time-sensitive (e.g., personality traits) is usually one or two weeks. However, previous research examining test-retest reliability of time-sensitive assessments such as mood outcomes advocated for a short interval between the assessments. The latter was deemed important to prevent day-to-day fluctuations affecting the reliability evaluation [16,17]. In the current study the assessments are time-sensitive, i.e. the outcomes can vary from day to day. Therefore, the two assessments (test and retest) were conducted on the same day. It was hypothesized that the single-item scales have a high test-retest reliability.

2. Materials and methods

The study was conducted in December 2021. On one test day, N = 108 participants (71.3% female, mean (SD) age of 21.5 (2.6 years old) completed two paper-pencil surveys. For test-retest assessments a sample size >100 is considered as excellent [18]. Participants were students of the department of pharmaceutical sciences of Utrecht University, The Netherlands. They were recruited via e-mail from a sample that previously participated in a study at Utrecht University [19], and via word of mouth. The study was reviewed and approved by the Science-Geo Ethics Review Board of Utrecht University (protocol ID: S-21525, date of approval: November 21, 2021), and all participants provided written informed consent. They received 20,- euro reimbursement for their participation. Participants were included if they were male or female, student, and between the age of 18–30 years old. There were no exclusion criteria. Given the considerable number of international students, participants could complete the survey in English or Dutch language. The same survey was completed twice by the same participants. The time between completion of the surveys was approximately 30 min. Participants were distracted by other tasks between the first and second survey. To further prevent participants from actively memorizing the answers provided to the first survey, they were not aware of the purpose of the second survey.

The assessed demographic data of the sample was limited to age and sex. Mood was assessed via 1-item scales including “stress”, “anxiety”, “depression”, “fatigue”, “loneliness”, “being active”, “optimism”, and “happiness”. All items were scored on a scale ranging from 0 (absent) to 10 (extreme) [2]. In a similar way, “quality of life” [2], and “immune fitness” [20–22] were assessed on a scale ranging from 0 (very poor) to 10 (excellent).

Statistical analyses were conducted with SPSS (IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 28. Armonk, NY, USA: IBM Corp.). Mean and standard deviation (SD) were computed for each variable. First, test and retest scores were compared with the Related-Samples Wilcoxon Signed Rank Test. Differences were considered significant if $p < 0.05$. Second, Pearson’s correlations were computed between the test and retest assessments. Correlations were considered statistically significant if $p < 0.05$. Third, to further evaluate reliability, intraclass correlations (ICC’s) were computed, and their 95% Confidence Interval (CI). A single-measurement, absolute-agreement, 2-way mixed-effects model was used to calculate the ICCs. To determine reliability, the 95% CI of the ICCs were interpreted as follows: 95% CI values less than 0.5 were considered indicative of poor reliability, values between 0.5 and 0.75 were considered indicating moderate reliability, values between 0.75 and 0.9 were considered indicating good reliability, and values greater than 0.90 were considered indicating excellent reliability [23]. Fourth, to confirm reliability, the Bland-Altman limits of agreement method was applied [24,25]. For each variable, the difference score (DIFF) of the test and retest outcomes and the corresponding standard deviation (SDDIFF) were computed. According to the limits of agreement method, there is agreement between the

Table 1
Test-retest assessments.

	Test	Retest	Correlation	
Variable	Mean (SD)	Mean (SD)	r	p-value
Stress	4.0 (2.7)	3.5 (2.5)	0.86	<0.001*
Anxiety	1.8 (2.2)	1.5 (2.0)	0.84	<0.001*
Depression	1.6 (2.0)	1.6 (2.0)	0.90	<0.001*
Fatigue	4.6 (2.4)	4.0 (2.4)	0.78	<0.001*
Loneliness	1.7 (2.2)	1.3 (1.9)	0.87	<0.001*
Optimism	6.3 (1.5)	6.2 (1.7)	0.70	<0.001*
Hostile	0.6 (1.4)	0.6 (1.4)	0.85	<0.001*
Happiness	6.6 (1.5)	6.6 (1.4)	0.80	<0.001*
Quality of life	7.5 (1.0)	7.4 (1.0)	0.67	<0.001*
Being active	6.2 (1.6)	6.0 (1.7)	0.73	<0.001*
Immune fitness	7.6 (1.2)	7.4 (1.3)	0.85	<0.001*

Notes: Pearson’s correlations were computed between the test and retest assessments. The correlations are considered statistically significant if $p < 0.05$, indicated by *.

assessments if 95% of the DIFF score lies between (DIFF - 1.96 x SDDIFF) and (DIFF + 1.96 x SDDIFF). No agreement between the test and retest assessment was concluded if 6% or more of the difference scores lie outside the limit of agreement interval.

3. Results

The mean (SD) of each variable, and the correlation between the test and retest session are summarized in Table 1. The Related-Samples Wilcoxon Signed Rank Test revealed no significant differences between the test and retest assessments. Significant Pearson's correlations ($p < 0.001$) were found between all test-retest assessments.

Intraclass correlations are listed in Table 2. The analysis confirmed that there was moderate to excellent agreement between the test and retest assessments.

Results of the Bland-Altman analysis are summarized in Table 3. It was predefined that no agreement would be concluded if 6% or more of the difference scores were outside the limits of agreement (LA) interval [24,25]. The analysis revealed agreement for all the assessments, except for depression which had a slightly higher percentage of participants that showed disagreement of the methods (6.5%).

4. Discussion

The analyses revealed a high test-retest reliability for single-item assessments of mood, immune fitness, and quality of life. These findings are important as they strengthen the rationale for using these single-item assessments.

In this study test-retest reliability was investigated using different approaches. However, it should be noted that the presented correlational analyses describe linear relationships between the test and retest assessments, but not whether the two assessments are in agreement [24,25]. A strong correlation between two assessments is no proof that they measure the same construct, i.e., that the assessments are in agreement. Highly significant correlations have been shown between variables that measure completely different constructs, such as bodyweight and height, or between alcohol consumption and smoking. Instead, to determine whether two assessments measure the same construct, Bland and Altman's the limits of agreement method is currently the gold standard [24,25]. Applying this method, agreement was found for all assessed single-items, except for depression. With respect to depression, it could be argued that the percentage of outcomes that were outside the agreement interval for this item was relatively low (6.5%), and in relation to the other assessments, its reliability could still be regarded as acceptable.

The use of single-item assessments is increasingly relevant in research and clinical practice. In clinical practice often real time information is needed for which traditional multiple item questionnaires are less suitable. Completing these elaborate questionnaires can be a burden for patients, whilst answering single-item questions is usually not considered effortful. For research purposes, the use of single-item assessments will significantly shorten surveys. They are also easily implemented in mobile phone apps, that become an increasingly important way of data collection [26,27]. Therefore, the current observation that single-item assessments are reliable is important.

While the use of single-item mood assessments has several advantages, there are also disadvantages that should be mentioned. The single item assessments provides a global measure, but no further details on the nature a construct or its impact on daily life. For example, traditional scales may have subscales (e.g., there are more types of anxiety, but also related constructs such as fear and worry may be measured within a multiple item scale). Also, considering individual items of a multiple item scale may provide more information on how a construct is affected or impacting daily life than using a single global measure. It therefore depends on the purpose of the investigation whether global single item assessments or multiple item scales are preferred.

Strengths of the current study were a sufficient sample size, and the fact that participants were unaware of the purpose of the study.

Table 2
Intraclass correlations between the test and retest assessment.

		95% CI		
Variable	ICC	Lower	Upper	Agreement
Stress	0.843	0.759	0.896	Good
Anxiety	0.824	0.747	0.878	Moderate to Good
Depression	0.899	0.856	0.930	Good to Excellent
Fatigue	0.759	0.635	0.839	Moderate to Good
Loneliness	0.844	0.755	0.898	Good
Optimism	0.689	0.575	0.776	Moderate to Good
Hostile	0.811	0.729	0.870	Moderate to Good
Happiness	0.796	0.715	0.856	Moderate to Good
Quality of life	0.672	0.555	0.764	Moderate to Good
Being active	0.724	0.620	0.803	Moderate to Good
Immune fitness	0.836	0.764	0.886	Good

Notes: To determine reliability, the 95% CI of the ICCs were interpreted as follows: 95% CI values less than 0.5 were considered indicative of poor reliability, values between 0.5 and 0.75 were considered indicating moderate reliability, values between 0.75 and 0.9 were considered indicating good reliability, and values greater than 0.90 were considered indicating excellent reliability [23]. Abbreviations: ICC = intraclass correlation, CI = confidence interval.

Table 3
Bland-Altman limits of agreement analysis.

Variable	Difference Mean (SD)	LA interval lower, upper	% outside the LA interval	Agreement
Stress	−0.49 (1.4)	−3.23, 2.25	5.6%	Agreement
Anxiety	−0.31 (1.2)	−2.66, 2.04	2.8%	Agreement
Depression	−0.06 (0.91)	−1.84, 1.72	6.5%	No agreement
Fatigue	−0.60 (1.6)	−3.74, 2.54	5.6%	Agreement
Loneliness	−0.41 (1.1)	−2.57, 1.75	2.8%	Agreement
Optimism	−0.13 (1.3)	−2.68, 2.42	5.6%	Agreement
Hostile	−0.14 (0.8)	−1.71, 1.43	5.2%	Agreement
Happiness	−0.56 (0.9)	−2.32, 1.20	5.6%	Agreement
Quality of life	−0.11 (0.8)	−1.68, 1.46	2.8%	Agreement
Being active	−0.21 (1.2)	−2.56, 2.14	3.7%	Agreement
Immune fitness	−0.18 (0.7)	−1.55, 1.19	4.6%	Agreement

Notes: No agreement is concluded if 6% or more of the difference scores are outside the limits of agreement (LA) interval [24,25].

That is, they were unaware that a retest session would take place and therefore it is very unlikely that they practiced and memorized the answers given in the first (test) survey. Limitations of the study may include that the surveys were administered in two languages. However, languages of the first and second survey were not mixed up within subjects. The use of single items also limited possible methodological issues related to text translation. Another limitation comprises the fact that mood can quickly fluctuate. This was the main reason to have the test and retest session on the same day, with only a short time interval of 30 min. However, also within 30 min mood can change. As a result, the assessments of the mood items do not perfectly fit (i.e., they closely correspond, but are not 100% the same for the test and retest session). In the current study, momentary assessments of mood were made. To overcome this issue in future studies, mood assessments could be made for a retrospective time period (e.g., past week). Finally, instead of random sampling, the study was conducted in a convenience sample of young adults (i.e., students). Therefore, future research should examine to what extent the current findings can be generalized to other age groups.

Taken together, the single-item assessments are ideal for screening purposes in clinical practice (e.g., identifying individuals at risk), or if one wishes a quick overall impression of a construct such as quality of life or stress. However, single-item assessments are less informative compared to multiple-item scales or extensive clinical interviews. Thus, in clinical practice it is recommended that poor scores on single-items are followed up with an interview with the patient.

5. Conclusion

The findings demonstrate the reliability of single-item assessments of mood, immune fitness, and quality of life.

Author contribution statement

Joris C Verster: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Kiki EW Mulder; Marjolijn CE Verheul; Evi C van Oostrom; Pauline A Hendriksen: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

Andrew Scholey; Johan Garssen: Conceived and designed the experiments; Wrote the paper.

Data availability statement

Data will be made available on request.

Funding

The authors received no specific funding for this work.

Declaration of competing interest

Over the past 3 years, J.C.V. has acted as a consultant/advisor for Eisai, KNMP, Red Bull, Sen-Jam Pharmaceutical, and Toast! J.G. is part-time employee of Nutricia Research and received research grants from Nutricia research foundation, Top Institute Pharma, Top Institute Food and Nutrition, GSK, STW, NWO, Friesland Campina, CCC, Raak-Pro, and EU. Over the past 3 years, A.S. has held research grants from Abbott Nutrition, Arla Foods, Bayer, BioRevive, DuPont, Kemin Foods, Nestlé, Nutricia-Danone, and Verdure Sciences. He has acted as a consultant/expert advisor to Bayer, Danone, Naturex, Nestlé, Pfizer, Sanofi, Sen-Jam Pharmaceutical, and has received travel/hospitality/speaker fees from Bayer, Sanofi, and Verdure Sciences. The other authors have no potential conflicts of interest to disclose.

References

[1] B.B. Hoeppner, J.F. Kelly, K.A. Urbanoski, V. Slaymaker, Comparative utility of a single-item versus multiple-item measure of self-efficacy in predicting relapse among young adults, J. Subst. Abuse Treat. 41 (2011) 305–312.

[2] J.C. Verster, E. Sandalova, J. Garssen, G. Bruce, The use of single-item ratings versus traditional multiple-item questionnaires to assess mood and health, Eur. J. Investig. Health Psychol. Educ. 11 (2021) 183–198.

[3] T. Baars, C. Berge, J. Garssen, J.C. Verster, Effect of raw milk consumption on perceived health, mood and immune functioning among US adults with a poor and normal health: a retrospective questionnaire based study, Compl. Ther. Med. 47 (2019), 102196.

[4] T. Baars, C. Berge, J. Garssen, J.C. Verster, The impact of raw fermented milk products on perceived health and mood among Dutch adults, Nutr. Food Sci. 49 (2019) 1195–1206.

[5] J.C. Verster, A. Anogeianaki, L.D. Kruisselbrink, C. Alford, A.-K. Stock, Relationship of alcohol hangover and physical endurance performance: walking the Samaria Gorge, J. Clin. Med. 9 (2020) E114.

[6] J.C. Verster, L. Arnoldy, A.J.A.E. van de Loo, A.D. Kraneveld, J. Garssen, A. Scholey, The impact of having a holiday or work in Fiji on perceived immune fitness, Tour. Hosp. 2 (2021) 95–112.

[7] J.C. Verster, L. Arnoldy, A.J.A.E. van de Loo, S. Benson, A. Scholey, A.-K. Stock, The impact of mood and subjective intoxication on hangover severity, J. Clin. Med. 9 (2020) 2462.

[8] J. Balikji, M.M. Hoogbergen, J. Garssen, J.C. Verster, Mental resilience, mood, and quality of life in young adults with self-reported impaired wound healing, Int. J. Environ. Res. Publ. Health 19 (2022) 2542.

[9] P.A. Hendriksen, P. Kiani, J. Garssen, G. Bruce, J.C. Verster, Living alone or together during lockdown: association with mood, immune fitness and experiencing COVID-19 symptoms, Psychol. Res. Behav. Manag. 14 (2021) 1947–1957.

[10] P.A. Hendriksen, J. Garssen, E.Y. Bijlsma, F. Engels, G. Bruce, J.C. Verster, COVID-19 lockdown-related changes in mood, health and academic functioning, Eur. J. Investig. Health Psychol. Edu. 11 (2021) 1440–1461.

[11] A. Merlo, N.R. Severeijns, S. Benson, A. Scholey, J. Garssen, G. Bruce, et al., Mood and changes in alcohol consumption in young adults during COVID-19 lockdown: a model explaining associations with perceived immune fitness and experiencing COVID-19 symptoms, Int. J. Environ. Res. Publ. Health 18 (2021), 10028.

[12] A.G. De Boer, J.J. van Lanschot, P.F. Stalmeier, J.W. van Sandick, J.B. Hulscher, J.C. de Haes, M.A. Sprangers, Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? Qual. Life Res. 13 (2004) 311–320.

[13] W.D. ′Killgore, The visual analogue mood scale: can a single-item scale accurately classify depressive mood state? Psychol. Rep. 85 (1999) 1238–1243.

[14] F. Wolfe, Fatigue assessments in rheumatoid arthritis: comparative performance of visual analog scales and longer fatigue questionnaires in 7760 patients, J. Rheumatol. 31 (2004) 1896–1902.

[15] J.C. Verster, A.J.A.E. van de Loo, S. Benson, A. Scholey, A.-K. Stock, The assessment of overall hangover severity, J. Clin. Med. 9 (2020) 786.

[16] R. Mesquita, D.J. Janssen, E.F. Wouters, J.M. Schols, F. Pitta, M.A. Spruit, Within-day test-retest reliability of the Timed up & Go test in patients with advanced chronic organ failure, Arch. Phys. Med. Rehabil. 94 (2013) 2131–2138.

[17] C.E. Paiva, E.M. Barroso, E.C. Carneseca, C. de Pádua Souza, F.T. Dos Santos, R.V. Mendoza López, et al., A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review, BMC Med. Res. Methodol. 14 (2014) 8.

[18] C.B. Terwee, L.B. Mokkink, D.L. Knol, R.W.J.G. Ostelo, L.M. Bouter, H.C.W. de Vet, Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist, Qual. Life Res. 21 (2012) 651–657.

[19] P.A. Hendriksen, A. Merlo, J. Garssen, E.Y. Bijlsma, F. Engels, G. Bruce, J.C. Verster, The impact of COVID-19 lockdown on academic functioning and mood: data from Dutch pharmacy students, PhD candidates and post-docs, Data 6 (2021) 120.

[20] L.J.F. Wilod Versprille, A.J.A.E. van de Loo, M. Mackus, L. Arnoldy, T.A.L. Sulzer, S.A. Vermeulen, et al., Development and validation of the immune status questionnaire (ISQ), Int. J. Environ. Res. Publ. Health 16 (2019) 4743.

[21] M. Van Schrojenstein Lantman, L.S. Otten, M. Mackus, D. de Kruijff, A.J.A.E. van de Loo, A.D. Kraneveld, et al., Mental resilience, perceived immune functioning, and health, J. Multidiscip. Healthc. 10 (2017) 107–112.

[22] J.C. Verster, A.D. Kraneveld, J. Garssen, The assessment of immune fitness, J. Clin. Med. 12 (2023) 22.

[23] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2016) 155–163.

[24] J.M. Bland, D.G. Altman, Statistical method for assessing agreement between two methods of clinical measurement, Lancet 327 (1986) 307–310.

[25] J.M. Bland, D.G. Altman, Measuring agreement in method comparison studies, Stat. Methods Med. Res. 8 (1999) 135–160.

[26] J.C. Verster, B. Tiplady, A. McKinney, Mobile technology and naturalistic study designs in addiction research, Curr. Drug Abuse Rev. 5 (2012) 169–171.

[27] J.C. Verster, A.J.A.E. van de Loo, S. Adams, A.-K. Stock, S. Benson, C. Alford, A. Scholey, G. Bruce, Advantages and limitations of naturalistic studydesigns and their implementation in alcohol hangover research, J. Clin. Med. 8 (2019) 2160.