Volume 38 / Number 1 / 2022

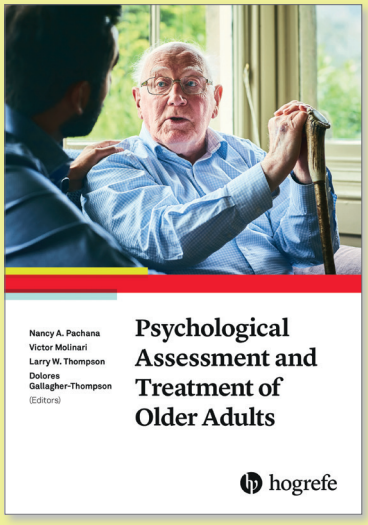
European Journal of Psychological Assessment

Official Organ of the European Association of Psychological Assessment

hogrefe

Expert guidance on working psychologically with older adults

New



Nancy A. Pachana / Victor Molinari / Larry W. Thompson / Dolores Gallagher-Thompson (Editors)

Psychological Assessment and Treatment of Older Adults

2021, viii/250 pp.
US $59.00 / € 50.95
ISBN 978-0-88937-571-0

Mental health practitioners are encountering an ever-growing number of older adults and so an up-to-date and comprehensive text addressing the special considerations that arise in the psychological assessment and treatment of this population is vital. This accessible handbook does just that by introducing the key topics that psychologists and other health professionals face when working with older adults. Each area is introduced and then the special considerations for older adults are explored, including specific ethical and healthcare system issues. The use of case examples brings the topics further to life.

An important feature of the book is the interweaving of diversity issues (culture, race, sexuality, etc.) within the text to lend an inclusive, contemporary insight into these important practice components. The Pikes Peak Geropsychology Knowledge and Skill Assessment Tool is included in an appendix so readers can test their knowledge, which will be helpful for those aiming for board certification in geropsychology (ABGERO).

This an ideal text for mental health professionals transitioning to work with older clients, for those wanting to improve their knowledge for their regular practice, and for trainees or young clinicians just starting out.

www.hogrefe.com

hogrefe

European Journal of

Psychological Assessment

Volume 38 / Number 1 / 2022

Official Organ of the European Association of Psychological Assessment

hogrefe

Founders

Rocío Fernández-Ballesteros and Fernando Silva

Supporting Organizations

The journal is the official organ of the *European Association of Psychological Assessment* (EAPA). The EAPA was founded to promote the practice and study of psychological assessment in Europe as well as to foster the exchange of information on this discipline around the world. Members of the EAPA receive the journal in the scope of their membership fees. Further, the Division for Psychological Assessment and Evaluation, Division 2, of the *International Association of Applied Psychology* (IAAP) is sponsoring the journal: Members of this association receive the journal at a special rate (see below).

Publisher

Hogrefe Publishing, Merkelstr. 3, D-37085 Göttingen, Germany, Tel. +49 551 999-500, Fax +49 551 999-50111, E-mail publishing@hogrefe.com, Web https://www.hogrefe.com
North America: Hogrefe Publishing Corp., 44 Merrimac St., Suite 207, Newburyport, MA 01950, USA, Tel. +1 978 255 3700, E-mail customerservice@hogrefe-publishing.com, Web https://www.hogrefe.com

Production

Regina Pinks-Freybott, Hogrefe Publishing, Merkelstr. 3, D-37085 Göttingen, Germany, Tel. +49 551 999-500, Fax +49 551 999-50111, E-mail production@hogrefe.com

Subscriptions

Hogrefe Publishing, Herbert-Quandt-Strasse 4, D-37081 Göttingen, Germany, Tel. +49 551 50688-900, Fax +49 551 50688-998, E-mail zeitschriftenvertrieb@hogrefe.de

Advertising/Inserts

Melanie Beck, Hogrefe Publishing, Merkelstr. 3, D-37085 Göttingen, Germany, Tel. +49 551 999-500, Fax +49 551 999-50111, E-mail marketing@hogrefe.com

ISSN

ISSN-L 1015-5759, ISSN-Print 1015-5759, ISSN-Online 2151-2426

Publication

Published in 6 issues per annual volume (new in 2017; 4 issues from 2004 to 2016)

Subscription Prices

Calendar year subscriptions only. Rates for 2021: Institutions – from US $483.00/€370.00 (print only; pricing for online access can be found in the journals catalog at hgf.io/journalscatalog); Individuals – US $264.00/€199.00 (print & online). Postage and handling – US $16.00/€12.00. Single copies: US $85.00/€66.50 + postage and handling. Special rates: IAAP/Colegio Oficial de Psicólogos members: €129.00, US $164.00 (+ €18.00, US $24.00 postage and handling); EAPA members: Included in membership

Payment

Payment may be made by check, international money order, or credit card, to Hogrefe Publishing, Merkelstr. 3, D-37085 Göttingen, Germany, or, for North American customers, to Hogrefe Publishing Corp., 44 Merrimac St., Suite 207, Newburyport, MA 01950, USA.

Electronic Full Text

The full text of the *European Journal of Psychological Assessment* is available online at https://econtent.hogrefe.com and in PsycARTICLES.

Abstracting/Indexing Services

The journal is abstracted/indexed in Current Contents / Social & Behavioral Sciences (CC/S&BS), Social Sciences Citation Index (SSCI), Social SciSearch, PsycINFO, Psychological Abstracts, PSYNDEX, ERIH, and Scopus. 2020 Impact Factor 2.985, 5-year Impact Factor 3.106, *Journal Citation Reports* (Clarivate Analytics, 2021)

Contents

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

https://doi.org/10.1027/1015-5759/a000699

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

Original Article

The Effects of Cognitive Load on Strategy Utilization in a Forced-Choice Recognition Memory Performance Validity Test

Elad Omer and Yoram Braw

Department of Psychology, Ariel University, Israel

Abstract: Despite the importance of detecting feigned cognitive impairment, we have a limited understanding of the theoretical foundation of the phenomenon and the factors that affect it. Studies regarding the formation and implementation of feigning strategies during neuropsychological assessments are numbered, though there are indications that they tax cognitive resources. The current study assessed the effect of cognitive load manipulation on feigning strategies. To achieve this aim, we utilized a 2 × 2 experimental design; condition (simulators/honest responders) and cognitive load (load/no load) were manipulated while participants (N = 154) performed a well-established performance validity test (PVT). The cognitive load manipulation reduced the quantity of feigning strategies, while also affecting their composition (i.e., strategies tended to be more intuitive). This suggests that reduced cognitive resources among those feigning cognitive impairment may impact the use of in-vivo feigning strategies. These findings, though preliminary, will hopefully encourage further research that will uncover the cognitive factors involved in the utilization of feigning strategies in neuropsychological assessments.

Keywords: feigned cognitive impairment, strategies, cognitive load, Performance Validity Test (PVT), Word Memory Test (WMT)

The feigning of cognitive impairment is a wide-spread phenomenon (Martin et al., 2015; Rogers & Bender, 2018) which can jeopardize the validity of neuropsychological assessments (Schutte et al., 2015). Consequently, the use of validity indicators, based either on stand-alone performance validity tests (PVTs) or standard cognitive tests, is currently considered a standard of practice (Bush et al., 2005; Heilbronner et al., 2009). This widespread use of validity indicators reflects its established efficacy (Lippa, 2018). Unfortunately, the focus on clinical utility has left a gap in our understanding of the cognitive underpinnings of feigned cognitive impairment, hindering progress in the field (Bigler, 2012; Eglit et al., 2017; Leighton et al., 2014).

Individuals who successfully feign brain injury employ systematic strategies that target aspects of test-taking performance (Kanser et al., 2017; Lau et al., 2017). For example, they may feign impairment only in tests that are thought to assess a specific cognitive domain (e.g., memory). However, despite their likely impact, we know comparatively little about the type of strategies that are utilized and the factors that affect them (as noted by Jones, 2017; Kanser et al., 2017). Regarding the latter, the formation and implementation of feigning strategies, as well as other deceptive behaviors, are likely affected by the examinee's cognitive resources (Vrij et al., 2017; Willison & Tombaugh, 2006). These, it is thought, can be depleted by various factors (e.g., the possibly anxiety-provoking nature of the assessment process). Consequently, these factors may lead to a less than optimal utilization of feigning strategies.

Cognitive load manipulations reduce the cognitive resources at the examinee's disposal (e.g., reducing working memory capacity; Ayres & Paas, 2012). For example, participants may be requested to perform a secondary distraction task while simultaneously performing the experiment's primary task, a commonly used cognitive load manipulation (Eglit et al., 2017). This methodology may therefore be an attractive method to achieve a better understanding of the interrelations between cognitive resources and strategy utilization (Driskell & Driskell, 2019). Two studies to date used the above-mentioned manipulation to investigate

feigned cognitive impairment. Batt et al. (2008) found that a dual-task interference paradigm significantly lowered the accuracy of individuals with severe brain injuries in one of the two forced-choice recognition memory tests (FCRM-PVTs) that were assessed (i.e., Word Memory Test; WMT), but not another (i.e., Test of Memory Malingering; TOMM). A later study of healthy adults similarly indicated the resistance of the TOMM to the effects of cognitive load manipulations (Barhon et al., 2015). Though innovative, these studies did not assess the effects of cognitive load manipulations on the use of feigning strategies. This is unfortunate, as such an investigation would enhance the theoretical understanding of feigned cognitive impairment and potentially advance daily clinical practices.

The aim of the current study was to assess the effects of a cognitive load manipulation on strategy utilization while feigning cognitive impairment. To achieve this aim, we used a 2 × 2 experimental design; simulators and honest controls performed a well-validated PVT (i.e., the WMT), either while concurrently performing a dual-task interference task or in a control condition (no-load). We hypothesized that the cognitive load manipulation would impact both the quantity and type of strategies used by simulators. First, it was hypothesized that manipulation would lower the number of strategies that they endorse. Second, studies suggest that the most frequently endorsed strategy by simulators is the feigning of memory impairment (i.e., deliberately lowering accuracy scores in tests that are perceived as assessing the examinee's memory). For example, the majority of simulators (76%) reported feigning memory loss as their primary feigning strategy in a study by Tan et al. (2002). In contrast, other strategies – such as slowing response times, feigning confusion, and poor concentration – were utilized by significantly fewer simulators. The latter strategies, it can be argued, are less intuitive and therefore likely tax cognitive resources. We therefore hypothesized that the cognitive load manipulation would decrease the simulators' propensity to use them. Finally, those feigning cognitive impairment may erroneously conclude that brain-injured patients perform delayed subtests of FCRM-PVTs more poorly than immediate recognition subtests. Correspondingly, the accuracy of simulators, but not controls, in Braw (2021) significantly decreased between the immediate recognition and delayed recognition subtests of the WMT (see also p. 37 in the WMT's manual; Green, 2005). We therefore hypothesized that simulators exposed to the cognitive load manipulation would exhibit a smaller decrease in accuracy between immediate and delayed subtests compared to those that were not exposed to the manipulation. Our hope was that this three-pronged approach would provide a glimpse of the possible effects of cognitive resources on examinees' performance in FCRM-PVTs and PVTs in general.

Method

Participants

Healthy adults participated in the study (N = 154). They were undergraduate students who received course credit for participating in the study and were excluded if they had any significant past/current neuropsychiatric disorders (age: 23.14 ± 2.86 years; education level: 12.40 ± 1.01 years). They were randomly divided into four groups (see Procedure section):

(1) Simulators exposed to cognitive load manipulation (SIM-LOAD; n = 38);
(2) Simulators not exposed to the manipulation (SIM-NoLOAD; n = 39);
(3) Honest controls exposed to cognitive load manipulation (HC-LOAD; n = 39); and
(4) Honest controls not exposed to the manipulation (HC-NoLOAD; n = 38).

For additional information, see Table 1. The study was approved by the University's Institutional Review Board (IRB) committee and all participants provided written informed consent before study entry.

Measures

Word Memory Test (WMT)

A computerized wordlist learning PVT (Green, 2003; Green et al., 1996). It includes six subtests, two of which are considered its primary performance validity subtests and were utilized in the current study: Immediate recognition (IR) and delayed recognition (DR). In the IR subtest, the examinees memorized a list of word pairs. Next, they were presented with word pairs, each containing a word that was previously presented and a foil. For each pair, they were requested to select the word which had appeared in the original list and were provided with feedback regarding the accuracy of their response. The DR subtest, identical to the IR subtest except for a change in foil words, was administered 30 min later (for an earlier study using the Hebrew version of the WMT, see Hegedish & Hoofien, 2013). Outcome measures in the current study were accuracy (% correct responses) and response time (RT_{Mean} in ms) for each subtest. An additional outcome measure (consistency, CNS) was generated by calculating the consistency of the participants' responses in the IR and DR subtests. See the WMT's manual for additional information (Green, 2005).

Table 1. Demographic and WMT outcome measures in the four experimental conditions (SIM-LOAD, SIM-NoLOAD, HC-LOAD, and HC-NoLOAD)

	Simulators (SIM)		Honest controls (HC)		
	Cognitive load (A)	No cognitive load (B)	Cognitive load (C)	No cognitive load (D)	
Measures	*n* = 38	*n* = 39	*n* = 39	*n* = 38	Statistical analyses
Demographic					
Age (years)	23.26 ± 3.33	23.79 ± 3.48	22.49 ± 2.30	23.00 ± 2.01	$F(3,150) = 1.42$ $p = .239$
Education (years)	12.55 ± 1.48	12.46 ± 0.91	12.36 ± 0.78	12.21 ± 0.70	$F(3,150) = 0.80$ $p = .497$
Gender (females)	33 (86.80%)	32 (82.10%)	34 (87.20%)	36 (94.70%)	$\chi^2(3) = 2.93$, $p = .403$
WMT's IR-subtest					
Accuracy (% correct)	44.87 ± 16.13	51.35 ± 18.63	91.67 ± 7.93	98.49 ± 3.37	–
RT_{Mean} (ms)	2,758 ± 2,601	3,413 ± 2,171	1,598 ± 499	1,238 ± 281	–
WMT's DR-subtest					
Accuracy (% correct)	44.93 ± 16.08	44.42 ± 15.71	91.02 ± 8.73	98.35 ± 3.69	–
RT_{Mean} (ms)	2,621 ± 2,598	2,471 ± 1,599	1,376 ± 435	1,029 ± 231	–
WMT general					
CNS (%)	60.20 ± 16.07	62.31 ± 12.41	90.00 ± 8.13	98.29 ± 3.49	–

Note. Data represent mean ± *SD*, except for gender. CNS = Consistency score; DR = Delayed recognition; HC = Honest controls; IR = Immediate recognition; RT_{Mean} = Mean response time; SIM = Simulators; WMT = Word Memory Test. For detailed statistical analyses of the WMT's outcome measures see Differences in Performance Between Immediate and Delayed Memory Subtests subsection in the Results section.

Strategy Utilization and Compliance Feedback Questionnaire

An open-ended self-report questionnaire that included two items: (a) *Strategy utilization*: The participants were requested to describe strategies that they utilized to feign cognitive impairment. More specifically, they were provided with the following instruction: "Please write down the strategies you used while performing the tests so as to present yourself as having a cognitive impairment". Participants were provided with an option to list up to three strategies and encouraged to be as specific as they could in their descriptions. (b) *Compliance*: Compliance with experimental instructions was assessed using a 7-point Likert scale, ranging from 1 to 7 (high scores indicating better compliance). While strategy utilization was measured solely among simulators, compliance was measured in both participant groups.

Dual-Task Interference

An auditory task that manipulates cognitive load (for studies using the manipulation, see Barhon et al., 2015; Batt et al., 2008; Craik, 1982). As part of the task, numbers (1–9) were presented to participants via an audio recording at 3-second intervals. Participants added the number three to each number presented to them, stating their answer aloud.

Procedure

Participants filled out a demographic-medical questionnaire and were then randomly assigned to one of four conditions, according to *group* (SIM/HC) and *cognitive load* (LOAD/NoLOAD). Simulators were provided with a script and requested to play the part of an examinee who has been involved in a car accident in which they sustained a concussion. Despite returning to their normal level of functioning, the examinee has been encouraged to falsify the existence of cognitive impairment and thereby receive a larger settlement in a lawsuit that was filed against their insurance company. Simulators were also requested not to present exaggerated impairments and to perform the task in a manner that would be deemed authentic, as recommended by Rogers and Bender (2018, pp. 592–614). Finally, they were notified that a cash prize of $100 would be raffled among the participants acting in accordance with the experimental instructions. In other words, simulators were provided with an incentive to succeed in feigning impairment while avoiding detection. In contrast, honest controls were requested to perform the tests to the best of their ability. To assess comprehension, participants were requested to describe the script and experimental instructions in their own words (i.e., the recall and comprehension elements of the recommended manipulation check; Rogers & Bender, 2018, pp. 599–600). Next, all participants performed the WMT's IR-subtest. Participants in the LOAD condition performed the dual-task interference for the duration of the subtest's learning phase, while those in the NoLOAD condition performed it without any interference. Between the administration of the IR and DR subtests, participants performed the CogState recommended battery for the assessment of traumatic brain injury (TBI) patients (i.e., Detection, Identification, One Card Learning, and One-Back tasks; Louey et al., 2014). After completion, participants filled out the feedback questionnaire. As detailed earlier, it included two items, the second of which assessed participants'

compliance. This item therefore constituted the second element of the manipulation check, termed "reported effort" by Rogers and Bender (2018, pp. 599–600).

Data Analysis

Preliminary Procedures

Response time outliers were identified based on individual RT distributions; response times faster than 200 ms and slower than 4 *SD*s above the individual mean were excluded from further analyses (similar to Braw, 2021; Schmiedek et al., 2007). Except for one participant who had two outliers and another who had three outliers, no more than one response time datum was discarded per participant. The groups were then compared in demographic variables using independent samples *t*-tests and χ^2 analyses (for parametric and non-parametric measures, respectively). For all comparisons, statistical significance was set at $p < .05$.

Analyses of Strategy Utilization

The use of strategies was analyzed using the following means:

Number of Endorsed Strategies

The total number of strategies used by each simulator was calculated based on the feedback questionnaire. The groups of simulators (i.e., SIM-LOAD/SIM-NoLOAD) were then compared using independent samples *t*-tests.

Type of Endorsed Strategy (Intuitive vs. Non-Intuitive)

The specific strategies that were noted in the feedback questionnaire were classified into broad categories using content analysis. This was accomplished in two consecutive stages. First, two research assistants separately classified the strategies into two broad categories; intuitive and non-intuitive strategies. The former included any reference to an attempt to manipulate accuracy measures, while the latter included a reference to all other feigning strategies. The research assistants were provided with representative examples (e.g., "I tried to make errors even when I knew the correct response") of the first category from the PIs (EO and YB). A third category included all cases in which no strategy was mentioned in the feedback questionnaire ("non"). A joint consultation was then performed among all research team members, leading to a unanimous agreement regarding the classification (i.e., regarding only one case was there an initial disagreement during the consultation). Next, both intuitive and non-intuitive strategies were subdivided in an iterative manner; the PIs performed a preliminary inspection of the feedback questionnaires and provided the two research assistants with subcategories based on common characteristics (e.g., "manipulation of response times," "inattentiveness," and "other" for the non-intuitive strategies). The research assistants then separately classified the intuitive and non-intuitive strategies and then conducted joint consultations with the PIs regarding each type of strategy. This was repeated twice, each time leading to a more refined list of subcategories (i.e., three and five subcategories for intuitive and non-intuitive strategies, respectively; see examples in the Electronic Supplementary Material, ESM 1). These subcategories were not overlapping, as each strategy fitted only one subcategory. A final joint consultation of all research team members was held with disagreements noted in only a few cases (5% of the strategies), cases in which agreement was reached unanimously during the consultation. Next, the number of endorsed strategies in each category (i.e., 0–3 intuitive and 0–5 non-intuitive) was calculated per simulator. The groups of simulators (SIM-LOAD/SIM-NoLOAD) were then compared using independent samples *t*-tests (α set at .025 following a Bonferroni correction).

Differences in Performance Between Immediate and Delayed Memory Subtests

Two repeated-measures analysis of variance (ANOVA) were used to analyze the WMT's outcome measures (i.e., accuracy and RT_{Mean}). Each included two between-subjects factors (*group*: SIM/HC; *cognitive load*: LOAD/NoLOAD) and a within-subject factor of *time* (IR subtest/DR subtest).

Exploratory Analysis

Each participant was scored as either passing or failing the WMT based on the WMT's classification scheme (Green, 2005, pp. 9–12). Next, the WMT's sensitivity and specificity were determined (Heilbronner et al., 2009, pp. 1119–1120). In addition, compliance of simulators and honest controls, based on the feedback questionnaire, was compared using an independent samples *t*-test.

Results

Analyses of Strategy Utilization

Number of Endorsed Strategies

Participants in the SIM-LOAD group reported using significantly fewer strategies (1.76 ± 0.75) than the SIM-NoLOAD group (2.18 ± 0.76), $t(75) = 2.42$, $p = .018$, $d = 0.55$.

Type of Endorsed Strategy (Intuitive vs. Non-Intuitive)

The proportions of participants reporting the use of an intuitive strategy were high in both groups (SIM-LOAD:

84.21%; SIM-NoLOAD: 74.36%). The groups also did not differ significantly in the endorsement of these intuitive strategies, $t(75) = -0.36$, $p = .723$, $d = 0.08$. This contrasted with the use of non-intuitive strategies. These were endorsed by markedly fewer simulators (i.e., 67.53%) and, importantly, SIM-LOAD participants reported using significantly fewer non-intuitive strategies (0.76 ± 0.82) than those belonging to the SIM-NoLOAD group (1.23 ± 0.74), $t(75) = 2.62$, $p = .010$, $d = 0.60$. Correspondingly, 52.63% of participants in the SIM-LOAD group reported using at least one non-intuitive strategy, compared to 82.05% in the SIM-NoLOAD group. Content analysis revealed that non-intuitive strategies included the manipulation of response times (slowing response times: 29.87%; impulsiveness: 22.08%), modifying performance according to the perceived level of difficulty (20.78%), inattentiveness/confusion (19.48%), and feigning based on assumptions regarding individuals with brain injury (7.79%). See Electronic Supplementary Material (ESM 1) for the description of strategies endorsed by the simulators, as well as illustrative quotes.

Differences in Performance Between Immediate and Delayed Memory Subtests

There were significant *group*, *cognitive load*, and *time* main effects in the ANOVA of participants' accuracy in the WMT's IR and DR subtests, $F(1, 150) = 643.39$, $p < .001$, $\eta^2 = .81$; $F(1, 150) = 6.92$, $p = .009$, $\eta^2 = .04$; $F(1, 150) = 7.29$, $p = .008$, $\eta^2 = .05$; respectively. Notably, the interaction effect was also significant, $F(1, 150) = 7.04$, $p = .009$, $\eta^2 = .04$. Post hoc paired-samples *t*-tests indicated that accuracy significantly decreased between the IR and DR subtests only among simulators that were not exposed to the cognitive load manipulation (i.e., SIM-NoLOAD), $t(38) = 3.22$, $p = .003$ (see Table 1 and Figure 1).

Regarding response times in the WMT, there were significant *group* and *time* main effects, $F(1, 150) = 35.16$, $p < .001$, $\eta^2 = .19$; $F(1, 150) = 30.75$, $p < .001$, $\eta^2 = .17$; respectively, while the *cognitive load* main effect was not significant, $F(1, 150) = 0.04$, $p = .843$, $\eta^2 < .001$. Notably, a significant interaction was evident, $F(1, 150) = 9.02$, $p = .003$, $\eta^2 = .06$. Post hoc paired-samples *t*-tests indicated a significant decrease in response times between the IR and DR subtests in all groups, except among simulators that were exposed to the cognitive load manipulation (SIM-LOAD), $t(37) = 1.10$, $p = .279$.

Exploratory Analyses

All simulators, irrespective of cognitive load condition, failed the WMT. The WMT's classification scheme was therefore associated with perfect sensitivity. However, its specificity was 89.6% since eight honest controls failed the WMT according to its classification scheme. Interestingly, significantly more controls in the load condition failed the WMT than those in the no-load condition ($n = 7$ vs. $n = 1$), $\chi^2(1) = 4.85$, $p = .028$, OR = 8.09. Specificity was therefore 82.0% in the HC-LOAD group and 97.4% in the HC-NoLOAD group. Regarding compliance with experimental instructions, no significant difference was found between simulators (5.96 ± 0.98) and honest controls (6.25 ± 1.04), $t(152) = 1.75$, $p = .081$, $d = 0.29$.

Discussion

Despite the importance of detecting feigned cognitive impairment, we have limited understanding regarding the choice and implementation of feigning strategies. By using a cognitive load manipulation, the current study modeled factors that reduced the examinees' available cognitive resources. More specifically, the rate and quality of feigning strategies were assessed while participants performed a well-established FCRM-PVT in either a cognitive load or a control condition.

Simulators reported the use of significantly fewer strategies when performing the WMT under cognitive load, compared to those who performed it in the control condition (SIM-LOAD and SIM-NoLOAD, respectively). A likely explanation is that the manipulation reduced their cognitive resources (Ayres & Paas, 2012) and consequently restricted the number of feigning strategies devised by the simulators. A theoretical explanation for this finding can be found in studies of other deceptive behaviors. These studies conceptualize most types of lying as cognitively more demanding than truth-telling (Depaulo et al., 2003; Vrij et al., 2001; Walczyk et al., 2005, 2018). Feigning cognitive impairment necessitates the examinee to inhibit correct responses, monitor and control their performance, and remind themselves throughout the assessment to act in a consistent and believable manner. As these factors have been suggested to reduce cognitive resources in other types of deceptions (Vrij et al., 2008), they likely also tax the cognitive resources of those feigning cognitive impairment. Correspondingly, neuroimaging studies suggest that falsifying memory requires greater activation of cognitive control networks (Kosheleva et al., 2016), while those that feign cognitive impairment have longer response times compared to presumably honest controls (see Braw, 2021). Regarding the latter point, honest examinees have relatively longer response times for incorrect than correct trials in FCRM-PVTs, while the RTs among simulators are comparable between the two types of trials (Kanser et al., 2019, p. 13). This suggests that response times of honest controls are slower only in the few items that they struggle to

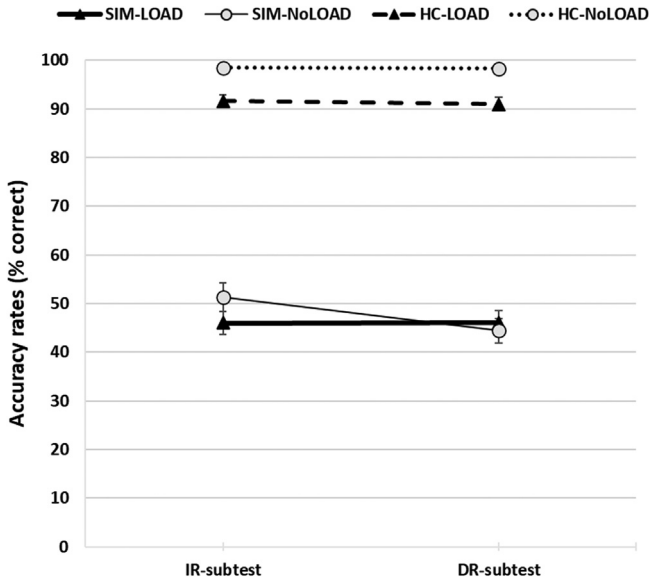


Figure 1. Accuracy rates of SIM-LOAD (n = 38), SIM-NoLOAD (n = 39), HC-LOAD (n = 39), and HC-NoLOAD (n = 38) groups in the WMT's IR and DR subtests. Error bars represent standard error of measurement (SEM). DR = Delayed recognition; IR = Immediate recognition; WMT = Word Memory Test.

remember, while the processing of those feigning cognitive impairment is longer for both types of responses. Though indirect, these findings have been interpreted to be the product of the additional cognitive processes that the deception entails and its cognitively taxing nature. Factors that *further* reduce the examinee's cognitive resources may consequently reduce the number of feigning strategies that are employed.

Deliberately making errors in FCRM-PVTs constitutes an intuitive feigning strategy utilized by most simulators. For example, purposeful incorrect responding and purposeful forgetting were the most frequently reported feigning strategies in Kanser et al. (2017). Similarly, Tan et al. (2002) found memory loss to be the most common strategy used by simulators (see also Iverson, 1995). Consistent with these earlier findings, the most frequently reported strategy by simulators in the current study was performing deliberate errors (79.22%), followed by less intuitive strategies (e.g., modifying performance according to the perceived level of task difficulty). Importantly, simulators in the cognitive load manipulation condition (SIM-LOAD) reported using significantly fewer non-intuitive feigning strategies than those in the control condition (SIM-NoLOAD), while the groups did not differ in the reported manipulation of accuracy rates (i.e., a relatively intuitive feigning strategy). Analyzing changes in participants' accuracy between the IR and DR WMT subtests mirrored the earlier reported findings. As indicated in a series of classic experiments, retention of briefly presented visual information, even over fairly long periods of time, is substantial (Nickerson, 1968). Examinees are therefore expected to maintain similar accuracy rates in the two WMT subtests. In contrast, those feigning cognitive impairment may erroneously presume that the memory trace fades with time and decrease their accuracy between immediate and delayed testing (for a description of detection methods based on the violation of learning principles, see Rogers & Bender, 2018). This was indeed found among simulators (SIM-NoLOAD) in the current study, as well as among outpatients that failed the WMT (Green, 2005, pp. 31–32). In contrast, simulators who were exposed to the cognitive load manipulation (SIM-LOAD) did not show this decrease in accuracy rates between subtests, suggesting that their reduced cognitive resources decreased their use of this feigning strategy. Furthermore, only these participants did not show a significant decrease in response times between the IR and DR subtests. Possibly, they continued to experience a reduction in their cognitive resources and did not benefit as other participants from prior exposure to the task (i.e., a practice effect). In summary, the cognitive load manipulation seemed to have a twofold effect on strategy endorsement, diminishing both the quantity and quality of feigning - strategies (i.e., fewer strategies were utilized and those that were used tended to be more intuitive). Interestingly, more honest controls in the load condition failed the WMT than those in the no-load condition. This decrease in specificity mirrors that found when individuals with brain injuries performed the WMT while concurrently performing an auditory distraction task (Batt et al., 2008). Though not the aim of the current study, the clinical implications of these findings should be explored in future studies.

Several limitations of the current study should be noted before summarizing its findings. First, simulation studies provide excellent control over internal validity and are therefore suited for exploratory research. That being said, a major limitation of this research design lies in its inability to ensure experimental instructions are appropriately followed by simulators and that supposedly honest controls actually perform the task to the best of their ability. This is always a concern, even when manipulation checks are in place, as in the current study. Moreover, the external validity of simulation research design has been criticized (Rogers & Bender, 2018, p. 11), including the daunting challenge to approximate actual incentives. To reduce the impact of this challenge, participants were given an incentive to succeed in presenting impairments in a manner that would be deemed credible. There is therefore a need for studies in real-life clinical settings to ensure the generalizability of the findings. Second, the current study used an open-ended self-report questionnaire to evaluate simulators' endorsed strategies. As it was filled out at the end of the experiment, the information gathered regarding strategies used while performing the WMT may have been distorted. Thus, it would be of interest to assess strategy utilization during the performance of the tests and not only at their completion, hopefully without interfering with test performance itself. Third, only one type of FCRM-PVT was employed in the current study. Researchers are encouraged to assess the effects of cognitive load manipulations on strategy utilization in other FCRM-PVTs, as well as standard neuropsychological tests. In addition to the suggestions listed earlier, researchers are encouraged to include coached simulators, especially using strategy-based coaching which has been proven as effective (Rogers & Bender, 2018, pp. 596–597), and to provide ample time and incentives for participants to devise strategies. This will enhance our understanding of the effects of cognitive load on feigning strategies that were devised prior to the start of the assessment, in contrast to the current study in which simulators devised them during the experimental session itself. Thereby, the ecological validity of these future studies will be enhanced. In addition, it would be of interest to assess the impact of cognitive load manipulations among neuropsychiatric patients. These studies may clarify whether these manipulations interact with the disorder to affect examinees' ability to form feigning strategies. Relatedly, these studies may pave the way to model, using cognitive load manipulation, the impact of certain neuropsychiatric conditions on feigning strategy formation and other cognitive processes of interest. Finally, the current study aimed to build a more solid foundation for clinical work and future theory-driven development of PVTs. Researchers can explore such clinical implications through the development of novel PVTs which are aimed at taxing participants' cognitive resources or, alternatively, by reducing such resources when performing existing PVTs (i.e., adding a cognitive load manipulation). This will potentially reduce strategy utilization and consequently allow easier identification of feigning. It should be noted, however, that this will necessitate a carefully planned research program. For example, increasing cognitive load will likely reduce specificity (i.e., more honest examinees may be erroneously labeled as feigners), stressing the need to evaluate the tradeoff between the positive and negative ramifications of each change in established procedures.

In conclusion, studies addressing strategy utilization of examinees feigning cognitive impairment are numbered. Applying cognitive approaches, such as inducing cognitive load, can enrich existing methods that aim to uncover cognitive processes involved in feigning behavior (Walczyk et al., 2018). The current study employed such an approach to investigate the effect of a cognitive load manipulation on strategy utilization, indicating that reduced cognitive resources impact both quantitative and qualitative aspects of in-vivo feigning strategies. Though hopefully providing an insightful theoretical glimpse into the phenomenon, these findings should be considered preliminary. Further studies are needed to both validate and extend the findings. Such studies may help uncover the cognitive factors involved in the formation of feigning strategies during neuropsychological assessments and potentially advance clinical practices.

Electronic Supplementary Materials

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1015-5759/a000636

ESM 1. Summary of strategies used by simulators.

References

Ayres, P., & Paas, F. (2012). Cognitive load theory: New directions and challenges. *Applied Cognitive Psychology, 26*(6), 827–832. https://doi.org/10.1002/acp.2882

Barhon, L. I., Batchelor, J., Meares, S., Chekaluk, E., & Shores, E. A. (2015). A comparison of the degree of effort involved in the TOMM and the ACS Word Choice Test using a Dual-Task Paradigm. *Applied Neuropsychology: Adult, 22*(2), 114–123. https://doi.org/10.1080/23279095.2013.863775

Batt, K., Shores, E. A., & Chekaluk, E. (2008). The effect of distraction on the Word Memory Test and Test of Memory Malingering performance in patients with a severe brain injury. *Journal of the International Neuropsychological Society, 14*(6), 1074–1080. https://doi.org/10.1017/S135561770808137X

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*(4), 632–642. https://doi.org/10.1017/S1355617712000252

Braw, Y. (2021). *Detection of feigned cognitive impairment: Utility of response time measurements in the performance validity subtests of the Word Memory Test.* Unpublished Manuscript.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology, 20*(4), 419–426. https://doi.org/10.1016/j.acn.2005.02.002

Craik, F. I. (1982). Selective changes in encoding as a function of reduced processing capacity. In F. Klix, J. Hoffmann, & E. van der Meers (Eds.), *Cognitive Research in Psychology* (pp. 152–161). North-Holland Publishing.

Depaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., & Charlton, K. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Driskell, T., & Driskell, J. E. (2019). Got theory? Multitasking, cognitive load, and deception. In T. Docan-Morgan (Ed.), *The Palgrave handbook of deceptive communication* (pp. 145–165). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-96334-1_8

Eglit, G. M. L., Lynch, J. K., & McCaffrey, R. J. (2017). Not all performance validity tests are created equal: The role of recollection and familiarity in the Test of Memory Malingering and Word Memory Test. *Journal of Clinical and Experimental Neuropsychology, 39*(2), 173–189. https://doi.org/10.1080/13803395.2016.1210573

Green, P. (2003). *Word Memory Test for Windows: User's manual and program*. Green's Publishing.

Green, P. (2005). *Green's word memory test for Microsoft Windows: User's manual* (rev. ed.). Green's Publishing.

Green, P., Allen, L., & Astner, K. (1996). *The Word Memory Test: A user's guide to the oral and computer-administered forms, US Version 1.1*. CogniSyst.

Hegedish, O., & Hoofien, D. (2013). Detection of malingered neurocognitive dysfunction among patients with acquired brain injuries. *European Journal of Psychological Assessment, 29*(4), 253–262. https://doi.org/10.1027/1015-5759/a000154

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *Clinical Neuropsychologist, 23*(7), 1093–1129. https://doi.org/10.1080/13854040903155063

Iverson, G. L. (1995). Qualitative aspects of malingered memory deficits. *Brain Injury, 9*(1), 35–40. https://doi.org/10.3109/02699059509004569

Jones, S. M. (2017). Dissimulation strategies on standard neuropsychological tests: A qualitative investigation. *Brain Injury, 31*(8), 1131–1141. https://doi.org/10.1080/02699052.2017.1283444

Kanser, R. J., Rapport, L. J., Bashem, J. R., Billings, N. M., Robin, A., Axelrod, B. N., & Miller, J. B. (2017). Strategies of successful and unsuccessful simulators coached to feign traumatic brain injury. *The Clinical Neuropsychologist, 31*(3), 644–653. https://doi.org/10.1080/13854046.2016.1278040

Kanser, R. J., Rapport, L. J., Bashem, J. R., & Hanks, R. A. (2019). Detecting malingering in traumatic brain injury: Combining response time with performance validity test accuracy. *The Clinical Neuropsychologist, 33*(1), 90–107. https://doi.org/10.1080/13854046.2018.1440006

Kosheleva, E., Spadoni, A. D., Strigo, I. A., Buchsbaum, M. S., & Simmons, A. N. (2016). Faking bad: The neural correlates of feigning memory impairment. *Neuropsychology, 30*(3), 377–384. https://doi.org/10.1037/neu0000251

Lau, L., Basso, M. R., Estevis, E., Miller, A., Whiteside, D. M., Combs, D., & Arentsen, T. J. (2017). Detecting coached neuropsychological dysfunction: A simulation experiment regarding mild traumatic brain injury. *The Clinical Neuropsychologist, 31*(8), 1412–1431. https://doi.org/10.1080/13854046. 2017.1318954

Leighton, A., Weinborn, M., & Maybery, M. (2014). Bridging the gap between neurocognitive processing theory and performance validity assessment among the cognitively impaired: A review and methodological approach. *Journal of the International Neuropsychological Society, 20*(9), 873–886. https://doi.org/10.1017/S135561771400085X

Lippa, S. M. (2018). Performance validity testing in neuropsychology: A clinical guide, critical review, and update on a rapidly evolving literature. *Clinical Neuropsychologist, 32*(3), 391–421. https://doi.org/10.1080/13854046.2017.1406146

Louey, A. G., Cromer, J. A., Schembri, A. J., Darby, D. G., Maruff, P., Makdissi, M., & Mccrory, P. (2014). Detecting cognitive impairment after concussion: sensitivity of change from baseline and normative data methods using the CogSport/Axon Cognitive Test Battery. *Archives of Clinical Neuropsychology, 29*(5), 432–441. https://doi.org/10.1093/arclin/acu020

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' Validity testing beliefs and practices: A survey of north american professionals. *The Clinical Neuropsychologist, 29*(6), 741–776. https://doi.org/10.1080/13854046.2015.1087597

Nickerson, R. S. (1968). A note on long-term recognition memory for pictorial material. *Psychonomic Science, 11*(2), 58–58. https://doi.org/10.3758/BF03330991

Rogers, R., & Bender, S. D. (2018). *Clinical assessment of malingering and deception*. Guilford Press.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136*(3), 414–429. https://doi.org/10.1037/0096-3445.136.3.414

Schutte, C., Axelrod, B. N., & Montoya, E. (2015). Making sure neuropsychological data are meaningful: Use of performance validity testing in medicolegal and clinical contexts. *Psychological Injury and Law, 8*(2), 100–105. https://doi.org/10.1007/s12207-015-9225-3

Tan, J. E., Slick, D. J., Strauss, E., & Hultsch, D. F. (2002). How'd they do it? Malingering strategies on symptom validity tests. *Clinical Neuropsychologist, 16*(4), 495–505. https://doi.org/10.1076/clin.16.4.495.13909

Vrij, A., Edward, K., & Bull, R. (2001). Stereotypical verbal and nonverbal responses while deceiving others. *Personality and Social Psychology Bulletin, 27*(7), 899–909. https://doi.org/10.1177/0146167201277012

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology, 22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior, 32*(3), 253–265. https://doi.org/10.1007/s10979-007-9103-y

Walczyk, J. J., Schwartz, J. P., Clifton, R., Adams, B., Wei, M. I. N., & Zha, P. (2005). Lying person-to-person about life events: A cognitive framework for lie detection. *Personnel Psychology, 58*, 141–170.

Walczyk, J. J., Sewell, N., & DiBenedetto, M. B. (2018). A review of approaches to detecting malingering in forensic contexts and promising cognitive load-inducing lie detection techniques. *Frontiers in Psychiatry, 9*(700), 1–14. https://doi.org/10.3389/fpsyt.2018.00700

Willison, J., & Tombaugh, T. N. (2006). Detecting simulation of attention deficits using reaction time tests. *Archives of Clinical Neuropsychology, 21*(1), 41–52. https://doi.org/10.1016/j.acn.2005.07.005

History
Received July 11, 2020
Revision received November 8, 2020
Accepted December 6, 2020
Published online March 2, 2021
EJPA Section / Category Cognitive Science

Acknowledgment
Special gratitude is given to the psychology undergraduate students, Eden Ofir-Katzav and Asaf Dubinsky, for aiding in the recruitment procedure of the participants and in collecting the data. Estie Arram Feder edited the final version of the manuscript and we thank her for this contribution to the paper.

Publication Ethics
The study was approved by the University's Institutional Review Board (IRB) committee and all participants provided written informed consent before study entry.

ORCID
Yoram Braw
https://orcid.org/0000-0002-5656-4863

Yoram Braw
Department of Psychology
Ariel University
POB 3
Ariel, 40700
Israel
yoramb@ariel.ac.il

Learn about new methods for investigating metamemory

Monika Undorf / Vered Halamish (Editors)

New Directions in Metamemory Research

Zeitschrift für Psychologie, vol. 228/4
2020, iv / 76 pp., large format
US $49.00 / € 34.95
ISBN 978-0-88937-581-9

Metamemory is ubiquitous in everyday life: We are confronted with what we know about our own learning and memory processes and how we assess and regulate these processes on a daily basis. The research collated here explores new methods for the study of metamemory (computational neuroimaging, verbal reports as data, and time-based measures) and addresses current and emerging research questions that advance our understanding of the basis, validity, and consequences of metamemory. It indicates links between metamemory and other related concepts, suggests how different perspectives on metamemory may be integrated, and identifies gaps in our knowledge about memory that deserve attention in future research. These new directions are made possible through the knowledge gained from diverse disciplines, including cognitive psychology, developmental psychology, educational psychology, and neuroscience.

Contents and topics include

- New Methods and Questions in Metamemory Research
- Toward a Neurocognitive Understanding of the Algorithms That Underlie Metamemory Judgments
- Time-Based Measures of Monitoring in Association With Executive Functions in Kindergarten Children
- Experience Matters: Effects of (In)Congruent Prompts About Word Frequency on Judgments of Learning
- Learning From (Test) Experience: Testing Without Feedback Promotes Metacognitive Sensitivity to Near-Perfect Recognition Memory
- Reactivity of Judgments of Learning in a Levels-of-Processing Paradigm
- The Language of Recollection in Support of Recognition Memory Decisions
- Louder = Larger = Clearer: Examining the Consistency of Metamemory Illusions

www.hogrefe.com

hogrefe

Original Article

You Can Play the Game Without Knowing the Rules – But You're Better Off Knowing Them

The Influence of Rule Knowledge on Figural Matrices Tests

Julie Levacher[1], Marco Koch[1], Johanna Hissbach[2], Frank M. Spinath[1], and Nicolas Becker[1]

[1]Department of Individual Differences & Psychodiagnostics, Saarland University, Saarbrucken, Germany
[2]Department of Biochemistry and Molecular Cell Biology, University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany

Abstract: Due to their high item difficulties and excellent psychometric properties, construction-based figural matrices tasks are of particular interest when it comes to high-stakes testing. An important prerequisite is that test preparation – which is likely to occur in this context – does not impair test fairness or item properties. The goal of this study was to provide initial evidence concerning the influence of test preparation. We administered test items to a sample of $N = 882$ participants divided into two groups, but only one group was given information about the rules employed in the test items. The probability of solving the items was significantly higher in the test preparation group than in the control group ($M = 0.61$, $SD = 0.19$ vs. $M = 0.41$, $SD = 0.25$; $t(54) = 3.42$, $p = .001$; $d = .92$). Nevertheless, a multigroup confirmatory factor analysis, as well as a differential item functioning analysis, indicated no differences between the item properties in the two groups. The results suggest that construction-based figural matrices are suitable in the context of high-stakes testing when all participants are provided with test preparation material so that test fairness is ensured.

Keywords: high-stakes testing, intelligence test, construction-based matrices test, test preparation, test fairness

It is largely acknowledged that intelligence is one of the best predictors of academic and professional success (e.g., Gottfredson, 2004; Roth et al., 2015; Schmidt & Hunter, 1998). In the context of high-stakes testing (e.g., student admission tests), figural matrices tests are promising because they are good indicators of general intelligence (*g*; e.g., Deary & Smith, 2004; Jensen, 1998; cf. Gignac, 2015 for a critical discussion), economical, and easy to administer. Several studies have indicated that two different strategies can be used to solve figural matrices (e.g., Bethell-Fox et al., 1984; Hayes et al., 2011; Jarosz & Wiley, 2012). Constructive matching consists of analyzing the rules applied in the matrix, cognitively generating the correct solution, and selecting it from the response format. Response elimination consists of comparing the response options with the item stem in order to eliminate as many distractors as possible and guessing from among the remaining ones. Response elimination can be seen as a fallback strategy that is used when respondents are not able to solve the items through constructive matching. By using response elimination, respondents with lower mental ability are able to enhance their test scores and show a performance that resembles that of respondents with higher mental ability. Recent research has demonstrated that this distortion reduces the item difficulty and convergent validity of classical figural matrices tests (Arendasy & Sommer, 2013; Becker, Schmitz, Falk, et al., 2016). A relatively new approach to solving this problem is the matrices construction task (Becker et al., 2015; Becker & Spinath, 2014), which requires participants to individually construct their own responses. Figure 1 presents an example of such an item. The item stem is the same as in classical figural matrices. It can be found in the upper part of the figure and consists of eight cells containing combinations of geometric figures that follow certain logical rules (e.g., the addition of symbols across the rows). The last cell is left empty, and respondents have to find the solution that completes the matrix according to the rules in the other cells. In contrast to classical matrices, the response choices in the lower part of the figure do not consist of predefined solutions but instead consist of a set of 24 symbols that have been used to construct the item stems. A respondent can thereby construct a complete response by choosing several of the appropriate symbols.

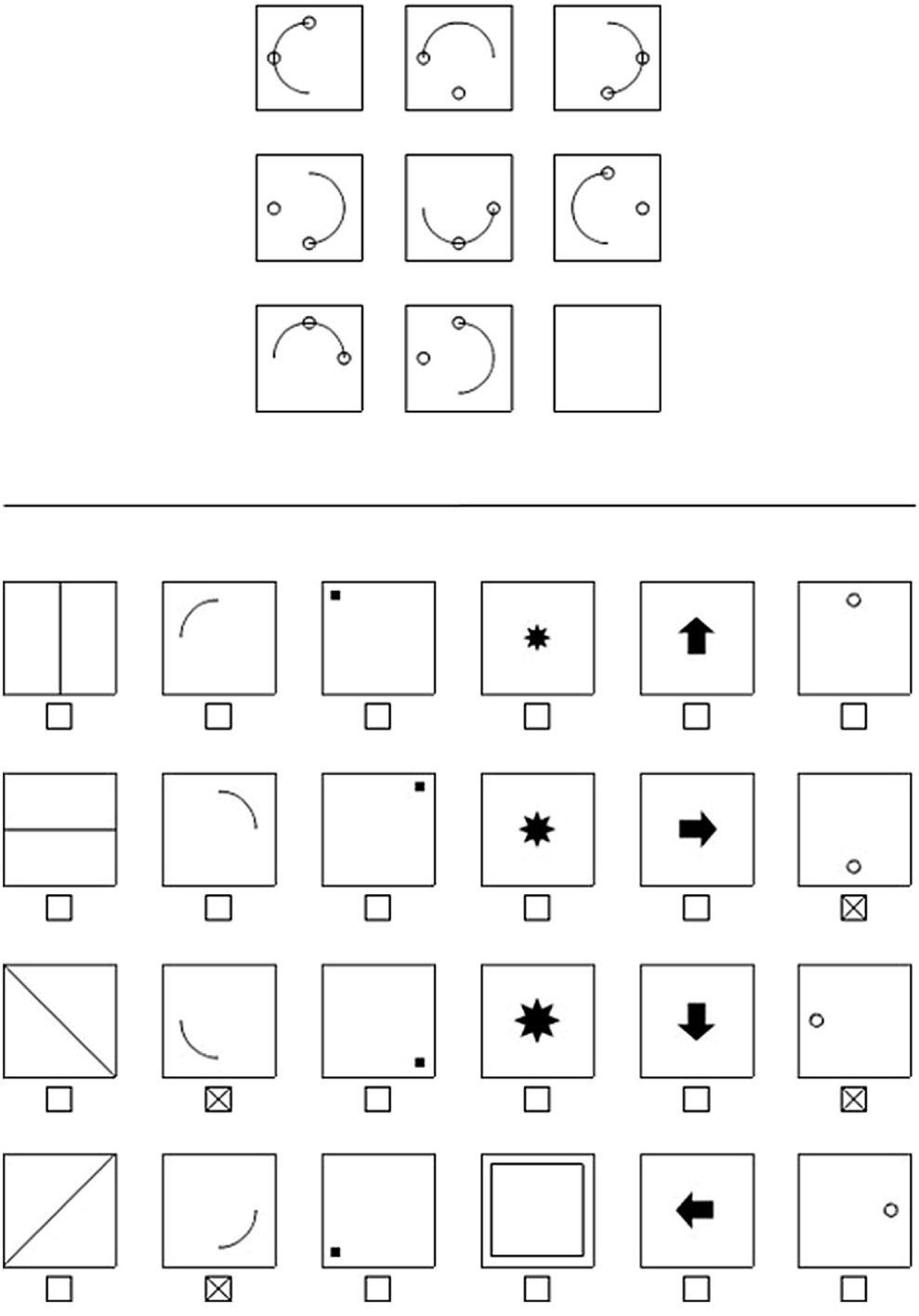


Figure 1. Example of an item stem and all possible answers. The correct answers are marked with an X.

In the context of high-stakes testing, test preparation is an issue that needs to be taken into account (Buchmann et al., 2010). Although slightly diverging categorizations of test preparation methods can be found in the literature (e.g., Arendasy et al., 2016; Hausknecht et al., 2007; Messick, 1982; Schneider et al., 2020), a differentiation between retesting, test familiarization, and test coaching seem plausible. Retesting consists of repeatedly taking the same test (at least twice) without receiving any further information about the correctness of the answers or specific test-taking strategies. Meta-analytic results have indicated that repeating a test increases test scores with a mean effect of $M(d) = .37$ (Scharfen et al., 2018). Test familiarization methods are aimed at increasing respondents' test-wiseness

by explaining formal aspects of the test (e.g., the number of items and time limits) and general test-taking strategies (e.g., how to use the answer sheet, time-management strategies) or by providing the opportunity to practice with alternate test forms. Test familiarization methods that are aimed at increasing test-wiseness have shown small to negligible influences on test results (Burns et al., 2008; Powers & Alderman, 1983). For test familiarization using alternate test forms, meta-analyses have reported mean effects of $M(d) = .23$ (Scharfen et al., 2018), $M(d) = .21$ (Hausknecht et al., 2007), and $M(d) = .23$ (Kulik et al., 1984). Test coaching is a more extensive intervention that – in addition to the contents of test familiarization – also provides information about the topics covered by the test and specific test-taking strategies as well as the opportunities to solve a larger set of items and to receive feedback on the chosen answer option (s). Meta-analytic results have indicated that coaching improves test scores with mean effects of $M(d) = .43$ (Kulik et al., 1984) and $M(d) = .64$ when combined with practice (Hausknecht et al., 2007). In the context of figural matrices, test coaching can consist of teaching respondents the rules employed in the item stems. Two recent studies compared respondents who watched a short video that explained the rules with unprepared control groups and found effects that were comparable or even better than the coaching effects reported in the meta-analyses (Loesche et al., 2015: $M(d) = .51$, $.14 \leq d \leq .81$; Schneider et al., 2020: $M(d) = 1.24$, $1.19 \leq d \leq 1.31$).

It is widely acknowledged that when scores increase as a result of test preparation, this does not represent gains in ability (Estrada et al., 2015; te Nijenhuis et al., 2007). Individual differences in test preparation might therefore mask ability differences and negatively affect test validity (cf. Messick, 1982). Schneider and colleagues (2020) conducted two studies to compare the correlations of the scores from a figural matrices test across different response formats with the scores from another intelligence test. Although they were comparably high in the test preparation ($r = .53/.58/.45$) and control groups ($r = .56/.53/.38$), in most cases, they were considerably lower when the two groups were combined ($r = .40/.42/.36$). This finding can be explained by the fact that respondents with comparable scores on the intelligence test achieved higher scores on the figural matrices test when they received test preparation materials. Loesche and colleagues (2015) compared the correlations of the scores from a figural matrices test and a working memory battery. In two of three experiments, the correlation in the test preparation group ($r = .62/.87/.61$) was higher than in the control group ($r = .46/.86/.42$). They speculated that these differences were caused by different solution strategies that differed in their demands on working memory. Unfortunately, they did not specify the corresponding correlations in the combined samples. Given the fact that the matrices test scores differed between the test preparation and control groups, it is, however, most likely that these results would be similar to Schneider and colleagues' (2020) results.

To conclude, the scores on cognitive ability tests in general and figural matrices tests, in particular, can be improved by test preparation. Taking into consideration that increases in scores due to test preparation do not represent gains in cognitive ability, convergent validity is threatened (and most likely criterion validity is too). Another problem is that tested fairness might be diminished because test preparation material is often expensive. Financially underprivileged respondents who cannot afford it might therefore suffer from disadvantages. Schneider and colleagues (2020) therefore suggested that all respondents be allowed to prepare in the same way in order to attenuate differences in rule knowledge and thereby improve the validity and fairness of figural matrices tests in high-stakes settings. With the current study, we present initial insights concerning brief test preparation material that simply consists of giving respondents written information about the rules employed in the test. Our first goal was to evaluate the effectiveness of the material and to compare it with the more extensive approaches used by Schneider and colleagues (2020) and Loesche and colleagues (2015). Additionally, we wanted to test whether the convergent validity concerning a natural science test and the criterion validity concerning scholastic achievement would diminish when combining groups of prepared and unprepared respondents. Following Loesche and colleagues' (2015) assumption of different solution strategies between the groups, we wanted to study the possible influences of test preparation on the item properties.

Methods

Sample and Procedure

The total sample consisted of 882 participants (71.09% female) with a mean age of 19.67 years ($SD = 2.01$, $16 \leq$ age ≤ 35) who were administered an admission test for a German medical school in August 2018. Participants were free to decide to take part in our study after they took the actual admission test but before they completed the whole day of testing. As compensation, they received feedback on their performance. The participants were randomly assigned to two different lecture halls in which the admission test took place. Participants in one lecture hall were given the test preparation material, whereas participants in the other lecture hall served as the control group. After providing informed consent, all participants received written instructions concerning the response format.

Table 1. Item difficulties and part-whole correlations for the test preparation group (p) and the control group (c)

Item	Rules	p_p	p_c	r_p	r_c	b_p	b_c	Wald	$Wald_p$	MH_Δ	MH_e
1	1	.94	.95	.31	.20	−2.35	−2.52	3.50	< .001	−1.79	C
2	1	.92	.90	.40	.34	−2.06	−1.92	2.79	.01	−1.25	B
3	1	.95	.93	.33	.24	−2.44	−2.29	1.95	.05	−0.69	A
4	1	.83	.80	.17	.12	−1.39	−1.26	4.65	< .001	−0.03	A
5	1	.84	.58	.45	.37	−1.41	−0.33	−2.70	.01	1.65	C
6	2	.63	.43	.43	.40	−0.47	0.26	1.46	.15	0.14	A
7	2	.77	.49	.45	.48	−1.04	0.03	−2.35	.02	1.46	B
8	2	.55	.32	.56	.43	−0.19	0.68	0.40	.69	−0.08	A
9	2	.67	.40	.57	.49	−0.60	0.36	−0.85	.39	0.29	A
10	2	.77	.51	.51	.58	−1.05	−0.06	−1.60	.11	0.70	A
11	2	.63	.46	.53	.36	−0.45	0.13	2.89	< .001	−0.62	A
12	2	.78	.63	.54	.47	−1.12	−0.48	1.31	.19	−0.50	A
13	2	.44	.21	.48	.40	0.20	1.18	−0.25	.81	0.23	A
14	3	.66	.51	.53	.54	−0.58	−0.06	3.33	< .001	−1.34	B
15	3	.60	.38	.59	.59	−0.36	0.45	0.85	.39	−0.86	A
16	3	.64	.37	.60	.58	−0.50	0.47	−0.84	.40	−0.08	A
17	3	.65	.33	.72	.62	−0.54	0.66	−2.89	< .001	0.15	A
18	3	.61	.42	.54	.54	−0.40	0.29	1.89	.06	−0.84	A
19	3	.57	.39	.51	.52	−0.25	0.41	2.36	.02	−0.83	A
20	3	.60	.23	.65	.57	−0.36	1.07	−4.70	< .001	1.45	B
21	3	.65	.36	.61	.63	−0.53	0.51	−1.52	.13	0.04	A
22	4	.39	.15	.52	.44	0.36	1.52	−1.52	.13	0.30	A
23	4	.45	.19	.61	.58	0.17	1.33	−1.75	.08	−0.45	A
24	4	.48	.20	.61	.49	0.08	1.28	−2.14	.03	0.23	A
25	4	.35	.11	.55	.51	0.53	1.83	−2.41	.02	0.22	A
26	5	.30	.11	.49	.46	0.73	1.90	−1.20	.23	−0.15	A
27	5	.35	.10	.56	.40	0.53	1.92	−2.96	< .001	0.72	A
28	5	.24	.07	.47	.45	0.97	2.29	−1.90	.06	−0.10	A

Notes. p_p = the probability that the test preparation group would solve the item; p_c = the probability that the control group would solve the item; r_p = part-whole correlation for the preparation group; r_c = part-whole correlation for the control group; b_p = item difficulty for the preparation group; b_c = item difficulty for the control group; Wald = *z*-statistic; $Wald_p$ = significance of the Waldtest; MH_Δ = Extent of the DIF according to the Mantel-Haenszel statistic; MH_e = Classifier for MH_Δ: A = no significant DIF; B = moderate DIF; C = large DIF.

In addition, the preparation group received detailed written information concerning the rules employed in the matrices test. The instructions used in the two groups can be found in the Electronic Supplementary Material, ESM 1. The 461 participants (69.63% female) in the test preparation group had a mean age of 19.70 years (SD = 2.08, 16 $\leq$ age $\leq$ 35). The control group consisted of 421 participants (72.68% female) with a mean age of 19.63 years (SD = 1.93, 17 $\leq$ age $\leq$ 33). Both groups had 10 min to read the instructions and 20 min to work on the test.

Matrices Test

The test used in this study consisted of 28 construction-based matrices. In accordance with Becker, Schmitz, Falk, and colleagues (2016), the items were constructed by applying six different rules (i.e., rotation, addition, completeness, subtraction, single element addition, intersection). Combinations of one to five rules were realized in the items (see Table 1).

External Criteria

We were able to match the results of the matrices tests with the results of a natural sciences test (Hamburg Natural Sciences Test; Ham-Nat; Meyer et al., 2019) that was part of the selection procedure. It captured knowledge of physics, chemistry, and biology. On average, the participants correctly solved M = 29.86 (SD = 8.75) of the 80 items. Furthermore, we had access to respondents' (GPA; in German "Abiturdurchschnittsnote") for which 1.0 was the best and 4.0 was the worst possible grade. The mean GPA in our study was M = 1.56 (SD = 0.19).

Statistical Procedure

Unless otherwise stated, all statistical analyses were carried out with the statistical software R (R Core Team, 2020). The R script used for this study and the resulting outputs can be found in ESM 2 and the syntax in ESM 3. We estimated McDonald's omega (ω) with the R package "coefficientalpha" (Zhang & Yuan, 2020) as well as the part-whole-corrected item-total correlations (r_{it}) in both groups to determine whether the reliability estimates would differ from one another. The item difficulties were calculated as the relative frequency with which participants solved the items as well as based on the one-parameter logistic model. To evaluate the influence of test preparation on the item difficulties, we used a paired-sample *t*-test. The latent mean test scores between the groups were compared with the R package "lavaan" (Rosseel, 2012) by fixing the latent mean of an arbitrary scale in one group to zero while letting the mean in the other group vary freely as described by Finch and French (2015). As an effect size index, we calculated Cohen's *d* with the R package "effsize" (Torchiano, 2016) and considered values on the order of $d \geq .20$, $d \geq .50$, and $d \geq .80$ as small, medium, and strong effects, respectively (Cohen, 1992). By computing correlations between the item difficulties, we tested whether they showed the same ranking in both groups. To determine differences in construct and criterion validity, we compared the correlations between the matrices test and the Ham-Nat as well as GPA in the two groups using the significance test provided by Diedenhofen and Musch (2015). A series of multigroup confirmatory factor analyses (MGCFAs) was estimated with the R package "lavaan" (Rosseel, 2012) in order to evaluate the measurement invariance of the matrices test in both groups. in accordance with Hirschfeld and von Brachel (2014), we conducted a series of confirmatory factor analyses of (1) a configural model with a single latent variable (i.e., *g*) and identical item properties between the two conditions; (2) a weak invariance model with equal loadings across the two groups; (3) a strong invariance model with equal loadings and intercepts across the groups; and (4) a strict invariance model with equal loadings, intercepts, and residual variances across the two conditions. Because the $\Delta\chi^2$ test is as sensitive to sample size as the χ^2 test (Tucker & Lewis 1973), differences between the models were assessed with the cut-offs suggested by Chen (2007), who proposed that ΔCFI < −.005 along with ΔRMSEA > .01 indicate non-invariance. To further examine the impact of the intervention on test fairness, we examined differential item functioning (DIF) in accordance with Penfield and Camilli (2006). One method to test for DIF is the Waldtest (carried out with the R package "eRm"; Mair & Hatzinger, 2007); however, since the Waldtest only highlights significant differences without any information about the extent of those differences, we also used the Mantel-Haenszel (carried out with the R package "difR"; Magis et al., 2010) statistic as it is more robust under less than optimal circumstances and offers effect sizes (Wetzel & Böhnke, 2017). Items were classified as having no significant DIF ($MH_\Delta < 1.0$), moderate DIF ($1.0 \leq MH_\Delta \leq 1.5$), or large DIF ($MH_\Delta > 1.5$; Magis et al., 2010).

Results

Reliability

The split-half reliability was comparably high in the two groups (test preparation group: r_{sh} = .88; control group: r_{sh} = .83). The same applied to the part-whole-corrected item-total correlations (test preparation group: *M* = 0.51, *SD* = 0.11; control group: *M* = 0.46, *SD* = 0.13). Additionally, the latent reliabilities were also excellent (test preparation group: ω = .93, 95% CI = [.91; .94]; control group: ω = .90, 95% CI = [.88; .92]).

Test and Item Difficulty

The difference in the latent means (ΔM = 1.48, *SD* = 1.49, 95% CI = [1.25; 1.71]) was significant [$\chi^2(1)$ = 154.09, $p < .001$] and indicated higher test scores in the test preparation group (*d* = 0.94, 95% CI = [0.80; 1.08]). Table 1 shows a comparison of the item difficulties, the difficulties determined in the one-parameter logistic model, the relative probabilities of solving the items, and the part-whole correlations for the test preparation and control groups. The item difficulties were significantly lower in the test preparation group [$t(54)$ = 3.06, p = .003, *d* = 0.82, 95% CI = [0.26; 1.38]; test preparation group: *M* = −0.52, *SD* = 0.87; control group: *M* = 0.34, *SD* = 1.22]. The probability of solving the items was higher in the test preparation group (*M* = 0.61, *SD* = 0.19) than in the control group (*M* = 0.41, *SD* = 0.25). The difference was significant [$t(54)$ = 3.42, p = .001] and had a strong effect (*d* = .92, 95% CI = [0.35; 1.48]). The correlation of the item difficulties between the test preparation and control groups (r = .99, $p < .001$) was strong and significant. Comparable correlations were found between the solution probabilities in the two groups (r = .94, $p < .001$).

Convergent and Criterion Validity

Results concerning convergent and criterion validity are presented in Table 2. The test scores in both groups were significantly correlated with the test score from the Ham-Nat (preparation group: r = .28, $p < .001$, control

Table 2. Correlations between the test scores and external criteria

Test scores	GPA	Ham-Nat
Total	−.06	.25*
Control group	.02	.24*
Test preparation group	−.08	.28*

Notes. GPA = grade point average; Ham-Nat = Hamburg Natural Sciences Test. *$p < .001$.

group: $r = .24$, $p < .001$), but the correlation coefficients did not differ significantly ($z = -0.71$, $p = .48$). The matrices test results were not significantly correlated with GPA in either group (test preparation group: $r = -.08$, $p = .07$, control group: $r = .01$, $p = .77$), and the difference between the two correlations was not significant ($z = -1.48$, $p = .14$). An examination of the correlations across the groups did not show meaningful differences from the groupwise analyses for the Ham-Nat score or for GPA.

Measurement Invariance

Table 3 shows the results for the MGCFA. Because an item-based model failed to converge, we employed the parceling strategy described by Matsunaga (2008). To this extent, we build four parcels, with the first parcel consisting of every fourth item beginning with the first item, the second parcel consisting of every fourth item beginning with the second item, and so on. The results indicated that strong measurement invariance could be assumed between the two groups. The ΔCFI remained below the threshold proposed by Chen (2007) for the weak and strong models. For the ΔRMSEA, the weak model exceeded the threshold, but for the strong model, it fell below the threshold. Both indices exceeded the threshold for the strict model. As Chen (2007) proposed that the ΔRMSEA should be used as a supplement to the ΔCFI, the results, therefore, indicated that the factor loadings and intercepts between the two groups were equal, whereas the residual variances were not. The results of the Waldtest and the Mantel-Haenszel statistic can be found in Table 1. Whereas the Waldtest was significant for 13 out of 28 items, the Mantel-Haenszel statistic indicated DIF for only six items. Only two items had an effect that indicated strong DIF, whereas the other four items were classified as having moderate DIF.

Discussion

With this study, we wanted to provide initial insights concerning brief test preparation material that consists of giving the respondents written information about the rules employed in a figural matrices test. Our results indicate that the respondents who received the test preparation material performed substantially better than the control group with an effect size ($d = .94$) that fell within the range of the effects found in previous studies (Loesche et al., 2015: $M(d) = .51$; Schneider et al., 2020: $M(d) = 1.24$). We, therefore, conclude that even rather minimal test preparation is sufficient for enhancing respondents' test scores. In contrast to the study by Schneider and colleagues (2020), combining the test preparation and the control groups did not lead to lower criterion or convergent validity. As the variance of GPA was substantially restricted ($M = 1.56$, $SD = 0.19$), the corresponding correlations could not be interpreted in a plausible way. Therefore, in future studies, it would be necessary to collect samples with higher variability in GPA. The correlations between the figural matrices test and the Ham-Nat did not indicate that the joint testing of prepared and unprepared respondents was problematic. A possible explanation of the differences concerning the study by Schneider and colleagues (2020) is that they used another intelligence test instead of a natural sciences test as we did. Differences due to test preparation might therefore be overshadowed by differences between the constructs that are considered when analyzing convergent validity. Future studies could investigate this assumption by employing different tests and testing whether the differences between the correlations in the separated and combined groups increase with the proximity of the constructs. Concerning the item properties, the MGCFA unfortunately did not converge on the level of single items. Although the results that were based on the item parcels have to be regarded as a simplification, they nevertheless suggest that the item properties did not change because the factor loadings and intercepts between the two groups were equal. Differences in the residual variances can be interpreted as representing individual differences in test preparation. This assumption was further strengthened by the results of the DIF analyses. Although the results of the Waldtest and the Mantel-Haenszel statistic differed with respect to the number of significant DIF effects between the two groups (13 vs. 6 out of 28 items), DIF was strong for only two items and moderate for four items. On the basis of an insightful comment offered by an anonymous reviewer, we conducted two post hoc analyses in which we correlated the number of rules and the differences in the solution probabilities and item difficulties between the two groups. The correlation between the difference in the solution probabilities and the number of rules was $r = .47$ ($p = .01$), and the correlation between the difference in the item difficulties and the number of rules was $r = .81$ ($p < .001$). These results indicate that with increasing numbers of rules, the difference in test performance between the groups increased. An interpretation for this finding might be that the facilitation of the solution process due to test preparation is particularly likely to affect complex items. Furthermore,

Table 3. Tests of measurement invariance

	Matrices with and without instruction									
Model	χ^2	*df*	$p(\chi^2)$	CFI	ΔCFI	RMSEA	ΔRMSEA	$\Delta\chi^2$	Δ*df*	$p(\Delta\chi^2)$
Configural	3.21	4	.523	1.00	–	< .001	–	–	–	–
Weak	12.22	7	.094	0.99	.001	.04	.04	9.01	3	.029
Strong	22.96	10	.011	0.99	.002	.05	.01	10.75	3	.013
Strict	177.05	11	< .001	0.94	.05	.19	.14	154.09	1	< .001

Notes. χ^2 = χ^2 value from the test of model fit; *df* = Degrees of freedom; $p(\chi^2)$ = Significance of the χ^2 value from the test of model fit; CFI = Comparative fit index; ΔCFI = Difference in CFI values between the model and the previous model; RMSEA = Root mean square error of approximation; ΔRMSEA = Difference in RMSEA values between the model and the previous model; $\Delta\chi^2$ = Difference in χ^2 values between the model and the previous model; Δ*df* = Difference in degrees of freedom between the model and the previous model; $p(\Delta\chi^2)$ = Significance of the difference in χ^2 values between the model and the previous model.

the near-perfect correlation between the item difficulties in the two groups indicated that changes in the mean difficulty did not influence the pattern of the item difficulties. We would therefore tend to assume that the item properties of the test are not influenced by test preparation. This provides additional support for Schneider and colleagues' (2020) suggestion to provide test preparation materials for all respondents in order to minimize the influence of such materials on the test results.

Apart from the aspects mentioned in the previous section, there are some limitations and perspectives for future research that need to be discussed. As already mentioned, the test preparation material employed in our study was rather short compared with corresponding interventions in previous studies (Loesche et al., 2015; Schneider et al., 2020). Although our results demonstrate that written information concerning the rules is sufficient for increasing test scores with an effect size comparable to previous studies, we do not know if more extensive test preparation would alter the item properties to a stronger extent. Future studies that compare test preparation materials with varying comprehensiveness are needed to answer this question. In this context, it would also be interesting to know if there is a saturation limit after which additional preparation efforts no longer pay off as this would be the minimal amount of preparation that should be provided to all candidates in order to ensure test fairness. Due to the fact that we did not use a computerized version of the test, we were not able to record meta-data (e.g., response times, the order in which the response options were selected), which might indicate behavioral differences between prepared and unprepared respondents. In the context of figural matrices tests, Becker, Schmitz, Göritz, et al. (2016) and Goldhammer et al. (2015) found that the relation between response time and item response is influenced by respondents' ability, and they interpreted this as an indicator of different solution strategies. Analyzing such differences between prepared and unprepared respondents could provide additional answers to the question of whether the solution strategies used to solve the test are influenced by test preparation.

To conclude, the results of our study demonstrate that test preparation is an issue that should be kept in mind when employing figural matrices tests in high-stakes testing. Differences in rule knowledge induced by test preparation can result in substantial differences in test performance that are independent of differences in ability. Although our results do not indicate changes in construct and criterion validity, we would not rule out that this is a potentially problematic issue. However, we suggest that test preparation most arguably does not change the item properties of figural matrices tests. In accordance with Schneider and colleagues (2020), we, therefore, argue that preparing respondents with written information concerning the rules employed in the items is a useful approach because it reduces individual differences in rule knowledge, and therefore, at the very least, it increases the fairness of the test.

Electronic Supplementary Materials

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1015-5759/a000637

ESM 1. Instructions used in the two groups

ESM 2. R script used for this study and the resulting outputs

ESM 3. R syntax without results as a RMD-file

References

Arendasy, M. E., & Sommer, M. (2013). Reducing response elimination strategies enhances the construct validity of figural matrices. *Intelligence, 41*(4), 234–243. https://doi.org/10.1016/j.intell.2013.03.006

Arendasy, M. E., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence, 55*, 44–56. https://doi.org/10.1016/j.intell.2016.01.004

Becker, N., Preckel, F., Karbach, J., Raffel, N., & Spinath, F. M. (2015). Die Matrizenkonstruktionsaufgabe: Validierung eines distraktorfreien Aufgabenformats zur Vorgabe figuraler Matrizen [The Construction Task: Validation of a distractor-free item format for the presentation of figural matrices]. *Diagnostica, 61*(1), 22–33. https://doi.org/10.1026/0012-1924/a000111

Becker, N., & Spinath, F. M. (2014). *Design a Matrix Test* (*DESIGMA*). Hogrefe.

Becker, N., Schmitz, F., Falk, A., Feldbrügge, J., Recktenwald, D., Wilhelm, O., Preckel, F., & Spinath, F. (2016). Preventing response elimination strategies improves the convergent validity of figural matrices. *Journal of Intelligence, 4*(1), Article 2. https://doi.org/10.3390/jintelligence4010002

Becker, N., Schmitz, F., Göritz, A., & Spinath, F. (2016). Sometimes more is better, and sometimes less is better: Task complexity moderates the response time accuracy correlation. *Journal of Intelligence, 4*(3), Article 11. https://doi.org/10.3390/jintelligence4030011

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8*(3), 205–238. https://doi.org/10.1016/0160-2896(84)90009-6

Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow education, American style: Test preparation, the SAT and college enrollment. *Social Forces, 89*(2), 435–461. https://doi.org/10.1353/sof.2010.0105

Burns, G. N., Siers, B. P., & Christiansen, N. D. (2008). Effects of providing pre-test information and preparation materials on applicant reactions to selection procedures. *International Journal of Selection and Assessment, 16*(1), 73–77. https://doi.org/10.1111/j.1468-2389.2008.00411.x

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Deary, I. J., & Smith, P. (2004). Intelligence research and assessment in the United Kingdom. In R. J. Sternberg (Ed.), *International handbook of intelligence* (pp. 1–48). Cambridge University Press.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One, 10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence, 50*, 93–99. https://doi.org/10.1016/j.intell.2015.02.004

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence, 52*, 71–79. https://doi.org/10.1016/j.intell.2015.07.006

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's Matrices. *Journal of Intelligence, 3*(1), 21–40. https://doi.org/10.3390/jintelligence3010021

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology, 86*(1), 174–199. https://doi.org/10.1037/0022-3514.86.1.174

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision, 11*(10), Article 10. https://doi.org/10.1167/11.10.10

Hirschfeld, G., & Von Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation, 19*(7), 1–12.

Jarosz, A. F., & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence, 40*(5), 427–438. https://doi.org/10.1016/j.intell.2012.06.001

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*(2), 179–188. https://doi.org/10.1037/0033-2909.95.2.179

Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven Advanced Progressive Matrices Test. *Intelligence, 48*, 58–75. https://doi.org/10.1016/j.intell.2014.10.004

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software, 20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures, 2*(4), 260–293. https://doi.org/10.1080/19312450802458935

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17*(2), 67–91. https://doi.org/10.1080/00461528209529246

Meyer, H., Zimmermann, S., Hissbach, J., Klusmann, D., & Hampe, W. (2019). Selection and academic success of medical students in Hamburg, Germany. *BMC Medical Education, 19*(1), Article 23. https://doi.org/10.1186/s12909-018-1443-4

Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 125–167) . Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X

Powers, D. E., & Alderman, D. L. (1983). Effects of test familiarization on SAT performance. *Journal of Educational Measurement, 20*(1), 71–79. https://doi.org/10.1111/j.1745-3984.1983.tb00191.x

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

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

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schneider, B., Becker, N., Krieger, F., Spinath, F. M., & Sparfeldt, J. R. (2020). Teaching the underlying rules of figural matrices in a short video increases test scores. *Intelligence, 82*, Article 101473. https://doi.org/10.1016/j.intell.2020.101473

te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence, 35*(3), 283–300. https://doi.org/10.1016/j.intell.2006.07.006

Torchiano, M. (2016). Effsize – A package for efficient effect size computation. *Zenodo*. https://doi.org/10.5281/ZENODO.1480624

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10. https://doi.org/10.1007/BF02291170

Wetzel, E., & Böhnke, J. R. (2017). Differential item functioning. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–5). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1297-1

Zhang, Z., & Yuan, K.-H. (2020). *coefficientalpha: Robust coefficient alpha and omega with missing and non-normal data*. R package version 0.7. https://CRAN.R-project.org/package=coefficientalpha

History
Received February 3, 2020
Revision received December 9, 2020
Accepted December 10, 2020
Published online March 2, 2021
EJPA Section / Category Intelligence

Julie Levacher
Differentielle Psychologie und Psychologische Diagnostik
Universität des Saarlandes
Campus A1 3
66123 Saarbrücken
Germany
julie.levacher@uni-saarland.de

Original Article

Psychometric Validation of a Parent-Reported Measure of Childhood Alexithymia

The Alexithymia Questionnaire for Children – Parent (AQC-P)

Ruth Harriet Brown[1], Aja Murray[1,2], Mary E. Stewart[3], and Bonnie Auyeung[1,4]

[1]Psychology Department, The University of Edinburgh, United Kingdom
[2]Institute of Criminology, University of Cambridge, United Kingdom
[3]Psychology, Heriot-Watt University, Edinburgh, United Kingdom
[4]Autism Research Centre, Department of Psychiatry, University of Cambridge, United Kingdom

Abstract: Alexithymia can be defined as difficulties in describing one's emotions and is of interest within clinical and developmental psychology as a potential mediating and exacerbating factor across multiple forms of psychopathology. Measuring alexithymia via self-reports can be challenging, as those with heightened alexithymia may have difficulties in recognizing their alexithymia traits due to impaired metacognitive skills. Thus, there would be considerable benefits to the availability of a psychometrically validated parent-reported alexithymia measure that may circumvent the issue of self-reports. We, therefore, examined the psychometric properties of a new parent-reported alexithymia measure, the Alexithymia Questionnaire for Children – Parent (AQC-P) in a community sample of 257 child-parent dyads. Furthermore, we examined the level of agreement between the parent-rated AQC-P and its self-rated counterpart, the Alexithymia Questionnaire for Children (AQC). Confirmatory factor analysis found that an oblique three-factor structure provided the best model for both AQC-P and AQC, with this structure showing measurement invariance across child gender. All subscales had omega internal consistency values > .70, supporting their reliability. Cross-informant consistency was supported by significant correlations between AQC and AQC-P scores. Results support the use of the AQC-P as a measure of parent-reported alexithymia in children.

Keywords: alexithymia, childhood, self-report, parent-report, psychometrics

Alexithymia refers to difficulties in identifying and describing one's emotions (Sifneos, 1973) and is thought to affect approximately 10% of the adult population (Mattila et al., 2006). While it is not itself recognized as a psychiatric illness (Ricciardi et al., 2015), it is considered to be a significant aggravating factor in many psychiatric conditions (Grabe et al., 2004) and is therefore of considerable interest to developmental and clinical psychologists. However, alexithymia measurement is challenging because those who would be expected to score high on these traits may lack the metacognitive skills required to recognize their alexithymic traits (Taylor et al., 1999). This issue may be further exacerbated in child populations due to their incomplete cognitive development and more limited abilities to introspect with their emotions (Parker et al., 2010). In this study we report on the validation of a new parent-reported measure of alexithymia to assess the construct in children, thereby potentially circumventing the issue of self-report.

Recognizing the possible limitations of self-reported measures of alexithymia, a small number of authors have previously developed observer-rated assessment tools for use in childhood yielding the Alexithymia Scale for Children – Teacher Form (ASC-TF; Fukunishi et al., 1998), the Children's Alexithymia Measure (CAM; Way et al., 2010), and the Alexithymia Questionnaire for Children – Parent (AQC-P; Costa et al., 2017). Only the CAM and the AQC-P were developed for use in parents, however, the former was previously shown to correlate nonsignificantly with the most widely used child measure of alexithymia, the Alexithymia Questionnaire for Children (AQC; Rieffe et al., 2006; see also Griffin et al., 2016. Further, only the AQC-P has items that correspond in content to the original adult alexithymia measure, the Toronto Alexithymia Scale-20 (TAS-20; Bagby et al., 1994), making it possible to

obtain a multi-informant perspective using comparable sets of items. However, the AQC-P has yet to undergo extensive psychometric validation, with only the measure's internal consistency reported in previous publications (Costa et al., 2017, 2019). Questions thus remain regarding its factorial validity, level of agreement with its self-reported counterpart the AQC, and gender invariance; the latter being important to test given previous evidence of a higher prevalence of alexithymia in males with the TAS-20 (Levant et al., 2009). Likewise, the AQC-P's convergent validity with known correlates of alexithymia (e.g., depressive symptoms; Parker et al., 1991, and empathic/prosocial behavior; Grynberg et al., 2010) remains unexplored. Thus, in this study, we provide an initial psychometric examination of the AQC-P in a community ascertained sample of 257 child-parent dyads.

Method

We report how we determined our sample size, all data exclusions, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact *p* values, effect sizes, and 95% confidence intervals (CIs).

Participants

Recruitment occurred over three waves of questionnaire distribution. First, children from Grade 4 to Grade 6 were recruited from two schools in the UK that agreed to participate; one local-authority and one private. In total, 521 families were approached to take part in the study. Of those approached, 175 families volunteered to take part, producing a response rate of 34%. Second, all the families in a university database of volunteers who were willing to take part in developmental studies and who had a child aged between 8 and 13 years (N = 45) were contacted 25 families of whom took part, producing a response rate of 55%. Finally, 57 families from a separate University of Edinburgh database of volunteers interested in taking part in research consented to complete the AQC-P. This gave an initial sample size of 257. Five children who had a diagnosis of autism spectrum disorder (ASD) were excluded from the sample to avoid confounding of ASD and alexithymic traits. This exclusion criterion was established prior to the data analysis.

The sample consisted of 121 boys and 131 girls between 8 and 13 years (M_{age} = 10.13; SD = 1.06), with 21.6%, 41.6%, and 36.8% from Grade 4, Grade 5, and Grade 6, respectively. The majority of parent participants were female (90%) and had a high level of education; 1.2% had no qualifications, 8.0% had school qualifications, 35.5% obtained an undergraduate degree, 26.4% obtained a postgraduate degree, and 11.2% obtained a doctoral degree. The remaining 17.6% (44) did not disclose their highest educational attainment. All participants were fluent in English.

Materials

All questionnaires were administered in English and as a child and parent booklet provided to participating families, containing a self-report questionnaire for children and a parent-report questionnaire for the parent(s).

Alexithymia Questionnaire for Children

The Alexithymia Questionnaire for Children (AQC; Rieffe et al., 2006) is a 20-item scale used to self-assess alexithymia in children as young as eight. It is built upon the three original subscales of the TAS-20 (Bagby et al., 1994); "difficulty identifying feelings" (DIF), "difficulty describing feelings" (DDF), and "externally oriented thinking" (EOT), where items are reworded to aid understanding by children. The total AQC has been previously found to have good internal consistency (α > .700; Rieffe et al., 2006). Using Rieffe and colleagues' (2006) scoring system, items were rated on a 3-point scale (1 = *not True* to 3 = *true*) in order to simplify the response scale for child participants. Scores ranged from 20 to 60, with higher scores indicating a greater degree of alexithymic traits.

Alexithymia Questionnaire for Children – Parent

The AQC was modified for use in the parent(s) of young children by Costa and colleagues (2017). The Alexithymia Questionnaire for Children– Parent (AQC-P) retains the same wording used by Rieffe and colleagues' (2006) (e.g., "I am able to describe my feelings easily" became, "my child is able to describe their feelings easily"), with the same three subscales DIF, DDF, and EOT. The measure has been previously found to have good internal consistency (α > .800; Costa et al., 2017). Like the AQC, the AQC-P is rated on a 3-point Likert scale in order to alleviate potential score comparison issues. Scores ranged from 20 to 60, with higher scores indicating a greater degree of alexithymic traits.

Depression Self-Rating Scale

The Depression Self-Rating Scale (DSRS; Birleson, 1981) is an 18-item self-reported scale that assesses depressive symptoms in children. The measure's scores were found to have good internal consistency in the current study (α = .948), similar to previous investigations (Birleson, 1981). Items are rated on a 3-point scale (0 = *never* to 2 = *always*). Scores

ranged from 0 to 36, with higher scores indicative of greater depressive symptoms.

Empathy Quotient for Children

The Empathy Quotient for Children (EQ-C; Auyeung et al., 2009) is a 27-item parent-reported assessment of empathic behavior in children. The measure's scores have been previously found to have good internal consistency (α > .900; Auyeung et al., 2009), similar to the current study (α = .856). Items were rated on a 4-point scale, with *definitely agree* and *slightly agree* responses endorsing empathic behavior scored as 2 and 1, respectively. *Definitely disagree* and *Slightly disagree* responses were scored as zero. Scores ranged from 0 to 54, with high scores indicative of higher empathy.

Strengths and Difficulties Questionnaire

The Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) is a 25-item parent-reported behavioral screening tool composed of five subscales that assess the child's negative and positive behavior; "conduct problems," "inattention-hyperactivity," "emotional symptoms," "peer problems," and "prosocial behavior." The SDQ scores were found to have good internal consistency (α = .722), similar to previous findings (Goodman, 1997). Items were rated on a 3-point scale (1 = *not true* to 3 = *certainly true*). Scores ranged from 5 to 15 for each subscale, with lower scores indicative of a behavioral strength and lower scores in the subscale "prosocial behavior" as a behavioral difficulty.

Procedure

For the school sample, questionnaire packs were distributed to pupils to be completed at home. For the database sample, participating families were sent their questionnaire packs via post. Before taking part, children and parent(s) were asked to read an information sheet and sign consent forms if they wished to participate. Families were informed that the study was voluntary and any information they provided would be anonymous. Parent(s) were additionally asked to provide demographic information on their child, including their age, gender, and any developmental difficulties. Both children and parents were asked to complete the questionnaire booklets independently, however, the children were informed they could ask their parent to clarify any item(s) that they did not understand. On completion, the questionnaires were either returned to school by the children to be collected by the researcher, sent back to the university via a pre-paid envelope, or given back to the researchers during the laboratory study. This study received ethical approval from the authors' research ethics committee (62-1516/6).

Data Analysis

Confirmatory factor analyses were conducted on the AQC and AQC-P scores in order to evaluate factorial validity. Three-factor structures were investigated. First, the original three-factor model (DIF, DDF, and EOT; Bagby et al., 1994) was explored. Next, a two-factor model proposed by Erni et al. (1997) (where DIF and DDF are merged together, known as "DDIF") has been suggested to be a more suitable specification for the TAS-20. Therefore, this specification was also explored. Lastly, for completeness, a one-factor model (combining DIF, DDF, and EOT) was investigated. Given the ordinal response format of the data, a diagonally weighted least squares (DWLS) estimation was used. Following the recommendations of Kline (2010), the suggested cut-offs < .060 for the root mean square error of approximation (RMSEA), < .080 for the standardized root mean square residual (SRMR), and > .900 for the comparative fit index (CFI) were used to identify adequate model-fit. Next, prior to investigating gender effects on AQC and AQC-P scores, measurement invariance was tested. As the data were categorical, the following was tested in the following order; configural invariance; threshold equivalence; threshold and loading invariance; and threshold, loading and intercept invariance. Invariance was assumed to hold if the differences in CFI, RMSEA, and SRMR between nested models were less than .005, .010, and .005, respectively (Chen, 2007). In order to assess internal consistency, McDonald's ω values were then calculated for the AQC and AQC-P total and subscales scores. Mean differences in alexithymia ratings between informants were then analyzed using paired *t*-tests. Cohen's *d* values were run in order to calculate effect sizes for these differences. Lastly, to assess convergent validity, correlations were calculated between the AQC, the AQC-P, and the additional questionnaires. Fisher *z*-transformations were used in order to identify any significant differences in the correlation coefficients. Analyses were conducted using both SPSS version 22, the R lavaan package for confirmatory factor analysis/measurement invariance (Rosseel, 2012), and psych package for the Fisher *z*-transformations (Revelle, 2018). Input/output files can be seen in the online deposits http://dx.doi.org/10.23668/psycharchives.4406 (supplementary file 1) and http://dx.doi.org/10.23668/psycharchives.4407 (supplementary file 2).

Results

Data Cleaning

Missing data were first addressed. Sixteen children and 15 parents omitted one or two items of their respective alexithymia measures. An expectation maximization (EM)

Table 1. Preliminary analysis

Measure	Mean	*SD*	α	Range	Skewness	Kurtosis
AQC-P	32.78	6.11	.822	20–51	.634	.025
AQC	35.51	5.75	.737	22–51	.136	−.591

Note. AQC = Alexithymia Questionnaire for Children; AQC-P = Alexithymia Questionnaire for Children – Parent.

Table 2. Confirmatory factor analyses of the one-, two- and three-factor solutions on the AQC-P and AQC

	χ^2	*df*	*p*	RMSEA	SRMR	CFI
AQC-P						
Model 1: 3 Factor	141.76	167	.922	.052	.054	.902
Model 2: 2 Factor	375.78	169	.000	.070	.075	.821
Model 3: 1 Factor	523.93	170	.000	.091	.091	.693
AQC						
Model 1: 3 Factor	235.53	167	.000	.041	.067	.963
Model 2: 2 Factor	316.17	169	.000	.059	.079	.830
Model 3: 1 Factor	341.63	170	.000	.064	.069	.801

Note. AQC = Alexithymia Questionnaire for Children; AQC-P = Alexithymia Questionnaire for Children – Parent; RMSEA = Root-Mean-Square Error of Approximation; SRMR = Standardized Root-Mean Square Residual; CFI = Comparative Fit Index.

algorithm within SPSS was used to estimate the values of the missing data points, allowing full scores to be generated. Two children were identified as multivariate outliers using Mahalanobis' distances. The child and their parent's data were removed from the dataset, giving a final sample size of 250 child-parent dyads.

Preliminary Analysis

Descriptive statistics for the AQC-P and AQC scores are shown in Table 1. The descriptive statistics for the external measures are shown in http://dx.doi.org/10.23668/psycharchives.4407 (supplementary file 3).

AQC-P Model Fit and Measurement Invariance

In order to investigate the fits of three-factor (model 1), two-factor (model 2), and one-factor (model 3) models, confirmatory factor analyses were conducted (see Table 2). The DWLS estimator produced a nonsignificant chi-square (χ^2) goodness of fit test for the three-factor model [$\chi^2(167) = 141.76$, $p = .922$] but not the two-factor model [$\chi^2(169) = 375.78$, $p < .001$] nor one-factor model [$\chi^2(190) = 523.93$, $p < .001$]. The criteria for adequacy of fit were met for the three-factor model, as satisfactory values for the CFI (> .900), RMSEA (< .060), SRMR (< .080) emerged. Despite an adequate SRMR value, no other goodness-of-fit tests were met in both the two-factor and one-factor models. Measurement invariance across gender was then tested in the three-factor model for the AQC-P. It was found ΔCFI, ΔRMSEA and ΔSRMR across the configural; threshold invariance; and threshold and loading invariance models were less than .005, .010, and .005, respectively (i.e., ΔCFI = .004, ΔRMSEA = .009, and ΔSRMR < .001) for the addition of threshold constraints; and ΔCFI = .004, ΔRMSEA = .001, and ΔSRMR < .001 for the addition of loading constraints), suggesting that invariance held up to the threshold and loading invariance level. However, ΔCFI was .007 and thus larger than .005, and ΔRMSEA was .012 and thus larger than .010 with the addition of intercept invariance constraints when added to the threshold and loading model (see http://dx.doi.org/10.23668/psycharchives.4407, supplementary file 4). Partial invariance up to the threshold, loading and intercept level (ΔCFI = .001, ΔRMSEA = .001 and ΔSRMR = .003) could be achieved when constraints were freed on item 20 (see http://dx.doi.org/10.23668/psycharchives.4407, supplementary file 5).

AQC Model Fit and Measurement Invariance

The above-described CFA and gender invariance analyses were also conducted for the AQC scores. First, the model fit of a one-, two- and three-factor structure was investigated (see Table 2). The DWLS estimator produced a significant χ^2 goodness of fit tests for the three-factor model [$\chi^2(167) = 235.53$, $p < .001$], the two-factor model [$\chi^2(169) = 316.17$, $p < .001$], and one-factor model [$\chi^2(170) = 341.63$, $p < .001$]. However, all other criteria for adequacy of fit were met for the three-factor model, as satisfactory values for the CFI (> .900), RMSEA (< .060), SRMR

(< .080) emerged from the analysis. While the one- and two-factor models produced satisfactory RMSEA values, CFI values were below acceptable levels for both models (< .900), and SRMR values were below acceptable levels for the one-factor model (< .060). Next, the degree of gender invariance was assessed using the three-factor model for the AQC (see http://dx.doi.org/10.23668/psycharchives.4407, supplementary file 6). While ΔRMSEA (.007) and ΔSRMR (< .001) were acceptable, ΔCFI was above the cut-off in the threshold invariance model (.006). However, follow-up analyses revealed partial threshold invariance (ΔCFI = .001; ΔRMSEA = .005; ΔSRMR = .002) when constraints were released on item 16 (see http://dx.doi.org/10.23668/psycharchives.4407, supplementary file 7). Furthermore, ΔCFI, ΔRMSEA, and ΔSRMR were less than .005, .010, and .005, respectively across the threshold and loading (ΔCFI $\leq$ .001; ΔRMSEA = .001; ΔSRMR $\leq$ .001) and threshold, loading and intercept invariance models (ΔCFI = .003; ΔRMSEA = .001; ΔSRMR = .001).

Reliability and Validity

Internal Consistency

Acceptable ω reliability values were found for the AQC-P total scores (ω = .870) and the DIF (ω = .890), DDF (ω = .700), and EOT (ω = .750) subscales. Likewise, the overall AQC produced acceptable ω values for the total scores (ω = .780) and the DIF (ω = .850), DDF (ω = .760) subscales. However, similar to the findings of Rieffe and colleagues (2006), the EOT subscale did not meet the acceptable level of internal consistency (ω = .560).

Correlations Between Total and Subscale Scores Across Raters

The total AQC and AQC-P scores were significantly correlated (r = .325, p < .001). At the subscale level, child and parent DIF (r = .401, p < .001), DDF (r = .206, p < .001), and EOT (r = .345, p < .001) all showed significant correlations.

Rating Differences Between the AQC and AQC-P

To asses if there were significant rating differences between the AQC and AQC-P scores, paired t-tests were conducted. Overall, children rated themselves more alexithymic than their parent, $t(249)$ = 6.26, p < .001, d = .461, 95% CI [1.87, 3.59]. At the subscale level, children gave higher DIF, $t(249)$ = 6.25, p < .001, d = .434, 95% CI [−1.67, −.870], and DDF, $t(249)$ = 7.11, p < .001, d = .573, 95% CI [−1.62, −.918], ratings. However, there was no significant difference between parent and child EOT scores, $t(249)$ = .954, p = .341, d = .071, 95% CI [−.598, .208].

Table 3. Correlations and Fisher z-transformations between the AQC, AQC-P and external measure scores

Measure	AQC	AQC-P	$z_{observed}$
DSRS	.628***	.291***	4.87***
EQ-C	−.281***	−.490***	2.75**
SDQ			
Prosocial behavior	−.239***	−.487***	3.21**
Inattention-hyperactivity	.204**	.382***	2.17*
Emotional symptoms	.193**	.296***	0.690
Conduct problems	.200**	.372***	2.09*
Peer problems	.066	.227***	0.610

Note. AQC = Alexithymia Questionnaire for Children; AQC-P = Alexithymia Questionnaire for Children – Parent; DSRS = Depression Self-Rating Scale; EQ-C = Empathy Quotient – Child; SDQ = Strengths and Difficulties Questionnaire. *p < .05; **p < .01; ***p < .001.

Convergent Validity

Fisher z-transformations suggested that the AQC-P correlated significantly stronger with the EQ-C ($z_{observed}$ = 2.75, p = .006); and the "prosocial behavior" ($z_{observed}$ = 3.21, p = .001), "inattention-hyperactivity" ($z_{observed}$ = 2.17, p = .030), and "conduct problems" (zobserved = 3.21, p = .036) subscales of the SDQ, compared to the AQC scores. Conversely, the AQC were found to correlate significantly stronger with DSRS scores ($z_{observed}$ = 4.87, p < .001) compared to the AQC-P (see Table 3).

Discussion

The aim of the current study was to assess the psychometric properties of the recently developed AQC-P to evaluate whether it is able to meet the current need for a parent-report measure of alexithymia, alongside the self-reported AQC. Analyses suggested that the AQC-P scores showed factorial validity with the instrument's hypothesized three-factor structure; good internal consistency; partial gender invariance up to the threshold, loading and intercept invariance level; significant correlations at the total score and subscale level with the AQC and convergent validity with the additional external measures administered.

A three-factor structure was found to be the best model for both the AQC-P and AQC scores as the models met all the goodness of fit tests, whereas one- and two-factor structures failed to reach the acceptable limits of model fit. These findings are consistent with previous investigations in the child (Rieffe et al., 2006) and adult (Taylor et al., 2003) samples using the AQC/TAS-20. A high level of partial measurement invariance was observed in both scales across boys and girls at the threshold, loading and intercept level after releasing constraints on the non-invariant items. As valid comparisons can still be drawn despite a small

number of non-invariant items (Pokropek et al., 2019), it was concluded that alexithymia was measured comparably across the genders with these scales. Thus, this supports the use of the AQC/AQC-P in examining gender differences in predictors/outcomes of alexithymic traits. Omega internal consistency values were all > .700 for the AQC-P total and subscale scores, though fell below .700 for the AQC EOT subscale.

At the total score and subscale level, the AQC and AQC-P scores were significantly correlated. Consistent with previous work in child psychopathology which suggests that self-and parent-reports capture overlapping yet distinct aspects of child behavior, the correlations were small to moderate in magnitude (De Los Reyes & Kazdin, 2005). Child self-reports yielded higher alexithymia ratings than parent reports. While it was not possible to gauge which informant provided the ratings that best reflected a child's true level of alexithymia, child-reports appeared to be more sensitive to detecting alexithymia than parent-reports.

However, Fisher *z*-transformations revealed that the AQC-P and AQC scores had unique correlation patterns with the external measures administered. Compared to the AQC, the AQC-P correlated more negatively with empathic and prosocial behavior. In contrast, the AQC was found to correlate more positively with depressive symptoms when compared to the AQC-P. Thus, while children can accurately report on their negative affect, they may have difficulties in reporting the external negative behaviors associated with alexithymia. In contrast, parents may accurately observe and rate their child's external negative behaviors, but may fail to detect their child's internal difficulties. Indeed, children rated themselves significantly higher on the DIF and DDF subscales when compared to their corresponding parent ratings. In comparison, no significant differences were observed between the child- and parent-rated EOT scores. Both instruments, while producing similar ratings, therefore appear to detect different degrees of cognitive and behavioral difficulties associated with alexithymia. Supporting this, post hoc analyses (see http://dx.doi.org/10.23668/psycharchives.4407, supplementary file 8) revealed that the AQC scores of the top 10% scoring children (n = 25) correlated nonsignificantly with any of the additional measures. In contrast, the corresponding AQC-P scores still correlated significantly negatively with empathic and prosocial behavior. Thus, the AQC-P appears to give a more accurate assessment of the child's associated difficulties when compared to the AQC. Despite this, further work is required to develop guidelines for combining the information from informants. For example, it is unclear whether higher scores based on single or multiple informants should be required to classify a child as at risk of alexithymia and, if the former, which informant's ratings best predict functional impairment. The most significant psychometric weakness identified was the reliability of the AQC EOT subscale. While the corresponding subscale in the parent-reported version has ω > .70, the child-reported EOT yielded an omega value of .560. Concerns have previously been raised regarding this subscale, with previous studies showing poor factor loadings and unacceptable internal consistency (see Bagby et al., 2020 for review). It may therefore be beneficial for future research to identify the core items of the EOT and revise or replace the items that have poor reliability.

Limitations

First, the sample was relatively small and recruited opportunistically. Bias may have been introduced as only families particularly motivated to take part in psychological experiments may have participated. Our preliminary results should thus be replicated in larger, more representative samples. Second, the age range of the sample was small and unequally distributed among the ages (i.e., 40% of the sample were 10-year-olds, whereas 1% were 13-year-olds). Consequently, child age measurement invariance could not be assessed. Future studies should assess the measurement invariance of the AQC and AQC-P's three-factor model in child populations with a more evenly distributed age range. The majority of parent reports were completed by the children's mothers. Collating psychometric assessment scores from both the child's mother and father has been recommended (Connell & Goodman, 2002). However, a high interrater agreement between mother- and father- reports of child behavior has been reported (Grietens et al., 2004), suggesting researchers can adequately rely on one parent to give an accurate evaluation of their child's behavior. Therefore, it is possible the data collected in the current study was not limited by the large proportion of mother informants. However, future investigations should investigate this by assessing the degree of cross-informant variance in mothers' and fathers' AQC-P scores. Concurrent validity of the AQC-P was not assessed as no additional parent-rated alexithymia was administered. Thus, it would be beneficial for future studies to assess the strength of the relationships between the AQC-P and other child-orientated observer-rated alexithymia scales (e.g., the CAM; Way et al., 2010). Lastly, future studies are required to translate the AQC-P into other languages, as the findings from the current study are only applicable to the English version.

Conclusions

Results support the factorial validity, reliability, convergent validity, gender invariance, and cross-informant correlations

of the AQC and AQC-P. This suggests that the recently developed AQC-P is a promising measure of parent-reported alexithymia that can be used alongside the AQC to provide a multi-informant measure of child alexithymia.

References

Auyeung, B., Wheelwright, S., Allison, C., Atkinson, M., Samarawickrema, N., & Baron-Cohen, S. (2009). The children's empathy quotient and systemizing quotient: Sex differences in typical development and in autism spectrum conditions. *Journal of Autism and Developmental Disorders, 39*(11), 1509–1521. https://doi.org/10.1007/s10803-009-0772-x

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The Twenty-Item Toronto Alexithymia Scale – I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research, 38*(1), 23–32. https://doi.org/10.1016/0022-3999(94)90005-1

Bagby, R. M., Parker, J. D., & Taylor, G. J. (2020). Twenty-five years with the 20-Item Toronto Alexithymia Scale. *Journal of Psychosomatic Research, 131*, Article 109940. https://doi.org/10.1016/j.jpsychores.2020.109940

Birleson, P. (1981). The validity of depressive disorder in childhood and the development of a self-rating scale: A research report. *Journal of Child Psychology and Psychiatry, 22*(1), 73–88. https://doi.org/10.1111/j.1469-7610.1981.tb00533

Brown, R. H., Murray, A. L., Stewart, M. E., & Auyeung, B. (2021a). *Code for: "Psychometric validation of a parent-reported measure of childhood alexithymia: The Alexithymia Questionnaire for Children – Parent (AQC-P)"*. https://doi.org/10.23668/psycharchives.4406

Brown, R. H., Murray, A. L., Stewart, M. E., & Auyeung, B. (2021b). *Supplementary materials for: "Psychometric validation of a parent-reported measure of childhood alexithymia: the Alexithymia Questionnaire for Children – Parent (AQC-P)"*. https://doi.org/10.23668/psycharchives.4407

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: a Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Connell, A. M., & Goodman, S. H. (2002). The association between psychopathology in fathers versus mothers and children's internalizing and externalizing behavior problems: A meta-analysis. *Psychological Bulletin, 128*(5), 746–773. https://doi.org/10.1037/0033-2909.128.5.746

Costa, A. P., Steffgen, G., & Samson, A. C. (2017). Expressive incoherence and alexithymia in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 47*(6), 1659–1672. https://doi.org/10.1007/s10803-017-3073-9

Costa, A. P., Steffgen, G., & Vögele, C. (2019). The role of alexithymia in parent–child interaction and in the emotional ability of children with Autism Spectrum Disorder. *Autism Research, 12*(3), 458–468. https://doi.org/10.1002/aur.2061

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*(4), 483–510. https://doi.org/10.1037/0033-2909.131.4.483

Erni, T., Lötscher, K., & Modestin, J. (1997). Two-factor solution of the 20-Item Toronto Alexithymia Scale confirmed. *Psychopathology, 30*(6), 335–340. https://doi.org/10.1159/000285079

Fukunishi, I., Yoshida, H., & Wogan, J. (1998). Development of the Alexithymia Scale for Children: A preliminary study. *Psychological Reports, 82*(1), 43–49. https://doi.org/10.2466/PR0.82.1.43-49

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Grabe, H. J., Spitzer, C., & Freyberger, H. J. (2004). Alexithymia and personality in relation to dimensions of psychopathology. *American Journal of Psychiatry, 161*(7), 1299–1301. https://doi.org/10.1176/appi.ajp.161.7.1299

Grietens, H., Onghena, P., Prinzie, P., Gadeyne, E., Van Assche, V., Ghesquiere, P., & Hellinckx, W. (2004). Comparison of mothers', fathers', and teachers' reports on problem behavior in 5-to 6-year-old children. *Journal of Psychopathology and Behavioral Assessment, 26*(2), 137–146. https://doi.org/10.1023/B:JOBA.0000013661.14995.59

Griffin, C., Lombardo, M. V., & Auyeung, B. (2016). Alexithymia in children with and without autism spectrum disorders. *Autism Research, 9*(7), 773–780. https://doi.org/10.1002/aur.1569

Grynberg, D., Luminet, O., Corneille, O., Grèzes, J., & Berthoz, S. (2010). Alexithymia in the interpersonal domain: A general deficit of empathy? *Personality and Individual Differences, 49*(8), 845–850. https://doi.org/10.1016/j.paid.2010.07.013

Kline, R. B. (2010). *Principles and practice of structural equation modelling*. Guilford Press.

Levant, R. F., Hall, R. J., Williams, C. M., & Hasan, N. T. (2009). Gender differences in alexithymia. *Psychology of Men & Masculinity, 10*(3), 190–204. https://doi.org/10.1037/a0015652

Mattila, A. K., Salminen, J. K., Nummi, T., & Joukamaa, M. (2006). Age is strongly associated with alexithymia in the general population. *Journal of Psychosomatic Research, 61*(5), 629–635. https://doi.org/10.1016/j.jpsychores.2006.04.013

Parker, J. D., Bagby, R. M., & Taylor, G. J. (1991). Alexithymia and depression: Distinct or overlapping constructs? *Comprehensive Psychiatry, 32*(5), 387–394. https://doi.org/10.1053/comp.2001.23147

Parker, J. D., Eastabrook, J. M., Keefer, K. V., & Wood, L. M. (2010). Can alexithymia be assessed in adolescents? Psychometric properties of the 20-Item Toronto Alexithymia Scale in younger, middle, and older adolescents. *Psychological Assessment, 22*(4), 798–808. https://doi.org/10.1037/a0020256

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling, 26*(5), 724–744. https://doi.org/10.1080/10705511.2018.1561293

Revelle, W. (2018). *psych: Procedures for personality and psychological research* (Version 1.8). http://CRAN.R-project.org/package=psych

Ricciardi, L., Demartini, B., Fotopoulou, A., & Edwards, M. J. (2015). Alexithymia in neurological disease: A review. *The Journal of Neuropsychiatry and Clinical Neurosciences, 27*(3), 179–187. https://doi.org/10.1176/appi.neuropsych.14070169

Rieffe, C., Oosterveld, P., & Terwogt, M. M. (2006). An Alexithymia Questionnaire for Children: Factorial and concurrent validation results. *Personality and Individual Differences, 40*(1), 123–133. https://doi.org/10.1016/j.paid.2005.05.013

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36.

Sifneos, P. E. (1973). The prevalence of "alexithymic" characteristics in psychosomatic patients. *Psychotherapy and Psychosomatics, 22*(6), 255–262. https://doi.org/10.1159/000286529

Taylor, G. J., Bagby, R. M., & Parker, J. D. (1999). *Disorders of affect regulation: Alexithymia in medical and psychiatric illness*. Cambridge University Press.

Taylor, G. J., Bagby, R. M., & Parker, J. D. (2003). The 20-Item Toronto Alexithymia Scale: IV. Reliability and factorial validity in different languages and cultures. *Journal of Psychosomatic Research, 55*(3), 277–283. https://doi.org/10.1016/S0022-3999(02)006018

Way, I. F., Applegate, B., Cai, X., Franck, L. K., Black-Pond, C., Yelsma, P., Roberts, E., Hyter, Y., & Muliett, M. (2010). Children's Alexithymia Measure (CAM): A new instrument for screening difficulties with emotional expression. *Journal of Child & Adolescent Trauma, 3*(4), 303–318. https://doi.org/10.1080/19361521.2011.609772

History
Received April 22, 2020
Revision received December 18, 2020
Accepted January 7, 2021
Published online April 22, 2021
EJPA Section / Category Differential Psychology

Acknowledgment
The authors wish to extend their gratitude to the families who participated in the study and to the reviewers who provided helpful feedback during the peer review process.

Publication Ethics
This study received ethical approval from the authors' research ethics committee (62-1516/6).

Open Science
Code: The R code and SPSS syntax for all analyses is available at http://dx.doi.org/10.23668/psycharchives.4406 (supplemental file 1) (Brown et al., 2021a).

Supplemental Materials: All other supplemental materials are available at http://dx.doi.org/10.23668/psycharchives.4407: 2 = Analysis output, 3 = Additional descriptive analyses, 4 = Measurement invariance of AQC-P results, 5 = Partial measurement invariance of AQC-P results, 6 = Measurement invariance of AQC results, 7 = Partial measurement invariance of AQC results, 8 = Post hoc analyses (Brown et al., 2021b).

Data: Data and materials will be made be available upon request.

Preregistration of Studies and Analysis Plans: This study was not preregistered and without an analysis plan.

Funding
This study received no grants or other financial support. Bonnie Auyeung was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skōōlodowska-Curie grant agreement No. 813546, the Baily Thomas Charitable Fund TRUST/VC/AC/SG/469207686, the Data Driven Innovation Programme, and the UK Economic and Social Research Council (ES/N018877/1) during the course of this work.

ORCID
Ruth H. Brown
https://orcid.org/0000-0002-1713-3585

Ruth Harriet Brown
Psychology Department
The University of Edinburgh
7 George Square
Edinburgh, EH8 9JZ
United Kingdom
rbrown11@exseed.ed.ac.uk

Brief Report

Does an Overall Job Crafting Dimension Exist?

A Multidimensional Item Response Theory Analysis

Leonidas A. Zampetakis

Applied Psychology Laboratory, Department of Psychology, University of Crete, Rethymno, Crete, Greece

Abstract: Job crafting is a multidimensional construct that can be conceptualized both at the general level and at the daily level. Several researchers have used aggregated scores across the dimensions of job crafting, to represent an overall job crafting construct. The purpose of the research presented herein is to investigate the factor structure of the general and daily versions of the job crafting scale developed by Petrou et al., (2012) (PJCS), using parametric multidimensional Item Response Theory (IRT) models. A sample of 675 employees working on different occupational sectors completed the Greek version of the scales. Results are in line with theoretical underpinnings and suggest that, although a bifactor IRT model offers an adequate fit, a correlated factors IRT model is more appropriate for both versions of the PJCS. Results caution against using aggregated scores across the dimensions of PJCS for both the general and daily versions.

Keywords: job crafting, multidimensional, Item Response Theory, bifactor, Greece

Job crafting is considered an important proactive approach to job redesign and is conceptualized as a multidimensional proactive behavior (Tims & Bakker, 2010). Systematic evidence from meta-analytic studies (Lichtenthaler & Fischbach, 2019; Rudolph et al., 2017) indicates that the dimensions of job crafting are interrelated, independent, not mutually exclusive, and have different antecedents and outcomes.

Several researchers, however, have used aggregated scores across the dimensions of job crafting to represent an overall job crafting construct (e.g., Akkermans & Tims, 2017), implying that job crafting is a general unidimensional factor. However, the use of aggregated scores across the dimensions to represent overall job crafting needs empirical justification.

In the present paper, we aim to investigate the factor structure of the general and daily level versions of the job crafting scale (PJCS) developed by Petrou et al. (2012) and modified from previous research (Tims et al., 2012). The PJCS consists of 13 items at the general level and 10 items at the daily level and differentiates between three types of job crafting behaviors, namely seeking resources, seeking challenges, and reducing demands.

We seek to contribute to a better understanding of the factorial structure of job crafting and make clear whether an overall factor of job crafting exists in the general and daily versions of the PJCS. Our research considers the multidimensional Item Response Theory (MIRT) as a viable approach to access the factor structure of PJCS. Compared to previous analyses of job crafting constructs (Bakker et al., 2018; Tims et al., 2012), which aimed at finding the smallest number of factors that reproduced the observed correlation matrix of the scales' items using factor analysis techniques, MIRT focuses in individual scale items and overcomes item-person confounding, found in classical test theory (CTT) techniques (like CFA). Moreover, in MIRT the ordinal level raw data of the rating scales (i.e., Likert scales), are transformed through logarithmic transformations into interval data and are not treated as continuous as is the case for CTT techniques (Reckase, 2009).

Method

Participants and Procedure

Data were based on the responses of 675 Greek employees (59.4% female), working in different occupational sectors (40% in the public sector). Participants were recruited through network sampling, from September 2019 to December 2019. The mean age was 39.80 years (SD = 10.87 years). The majority of the sample had a university degree (48.7%) and worked on average of 38.34 hrs per week (SD = 11.09 hrs).

https://doi.org/10.1027/1015-5759/a000638

Table 1. Cronbach's coefficient α, first eigenvalue, percentage of variance accounted for by the first eigenvalue, KMO, and Bartlett's statistic with degrees of freedom (*df*) from EFA analyses

Scale	Number of items	Cronbach's coefficient α	First eigenvalue	Percentage of variance explained	Kaiser-Meyer-Olkin (KMO) test	Bartlett's statistic (*df*)
General job crafting	13	0.76	3.65	28.16	0.788	2,854.9 (78)*
gSR	6	0.76	2.81	46.95	0.803	978.7 (15)*
gSC	3	0.86	2.35	78.41	0.698	1,051.0 (3)*
gRD	4	0.74	2.25	56.35	0.737	6,08.1 (6)*
Daily job crafting	10	0.75	3.48	34.87	0.765	3,076.5 (45)*
dSR	4	0.75	2.30	57.42	0.725	6,83.2 (6)*
dSC	3	0.90	2.54	84.74	0.749	1,380.2 (3)*
dRD	3	0.82	2.21	73.93	0.683	764.6 (3)*

Note. gSR = general seeking resources; gSC = general seeking challenges; gRD = general reducing demands; dSR = daily seeking resources; dSC = daily seeking challenges; dRD = daily reducing demands. *$p < .001$.

Measurement of Theoretical Constructs

Native speakers translated all the items of the main constructs used in the study, into the Greek language. A back-translation into English by other bilingual individuals revealed that the translation had worked quite well. All constructs included in the analysis were assessed with self-report measures. Responses to items were made on 5-point Likert scales. Two different forms of the questionnaires were presented to respondents in order to counterbalance the order of the job crafting constructs.

General Job Crafting

We adopted the three scales of general job crafting used by Petrou et al. (2012). Respondents were asked to indicate how often they engage in several behaviors in general. Table 1 presents Cronbach's reliability coefficients for the three scales.

Daily Job Crafting

We adopted the three scales used by Petrou et al. (2012) for the daily version of job crafting. Respondents were asked to indicate how often they engaged in several behaviors during the past day. Table 1 presents Cronbach's reliability coefficients for the three scales.

Statistical Procedure

We examined whether data from both scales are "unidimensional enough" (i.e., in exploratory factor analysis [EFA] the first factor accounts for at least 20% of the total variance; Anderson et al., 2017). We used EFA (FACTOR v.10.8 program) and the minimum rank factor analysis (MRFA) with direct Oblimin rotation for factor extraction. Results of EFA, in Table 1, support the existence of sufficiently unidimensional factors.

Then, we examined the fit of four unidimensional polytomous IRT models: The Partial Credit Model (PCM), the Generalized Partial Credit Model (GPCM), the Rating Scale Model (RSM), and the Graded Response Model (GRM) (for equations of models see De Ayala, 2013). To determine the model with the best fit, we used two indices across models: the Bayesian information criterion (BIC) and Akaike's information criterion (AIC).

Next, we fitted multidimensional alternatives of the unidimensional IRT model with the best fit. We examined model-data fit and item fit, using several statistical procedures. (1) M_2 goodness-of-fit statistic and the associated root mean square error of approximation (RMSEA). For good model-fit, the M_2 statistic is not significant ($p > .05$), and RMSEA with values close to zero to indicate an acceptable model fit. (2) Standardized local dependence (LD) χ^2, with values greater than |10|, indicating likely LD. (3) S-χ^2 item-fit diagnostic statistic, with significant values indicating lack of fit. For the analyses, we used the computer program IRTPRO (v.4.20) (Cai et al., 2017).

Results

Classical Item Analysis of the PJCS

Results, of the classical item analysis of the PJCS, are presented in Table E1 of the Electronic Supplementary Material, ESM 1. The indices indicate that most of the items have good discrimination. Items of the reducing demands subscale have relatively low discriminations (below 0.40). The mean item scores indicate that most of the items are "easy" in that the mean is above the midpoint on the scale. We found positive and statistically significant correlations between subscale scores for the general ($r_{gSR,\ gSC} = 0.45$, $p < .001$; $r_{gSR,\ gRD} = 0.04$, $p = .19$; $r_{gSC,\ gRD} = 0.13$, $p < 0.001$) and daily version ($r_{dSR,\ dSC} = 0.50$, $p < .001$; $r_{dSR,\ dRD} = 0.12$, $p < .001$; $r_{dSC,\ dRD} = 0.01$, $p = .77$) of the PJCS.

Unidimensional IRT Model Comparisons

Results of the model selection procedure (see Table E2 of ESM 1) for the unidimensional IRT models suggests that the GRM could be selected as the best model for the two overall job crafting scales.

Evaluation of Local Dependence

Results of the standardized LD-χ^2 with values larger than 10 suggested that item responses covariation in both scales maybe better modeled by a multidimensional model such as the multidimensional GRM (mGRM) or the bifactor GRM (bGRM). We fitted a multidimensional GRM and a bifactor GRM to both versions of the overall job crafting scale (see Tables E3a and E3b of ESM 1).

Global Model-Data Fit and Comparison for the GRM, mGRM, and bGRM

The bottom half of Table E4 of ESM 1, displays a summary of the model comparison and fit results for GRM, mGRM, and bGRM. All indexes (−2LL, BIC, AIC) agree that the best fitting model is the bifactor model for both versions of the overall job crafting scale.

Parameter Estimates for the bGRM

Tables E5 and E6 of ESM 1 summarize the bGRM parameter estimates for the general and daily overall versions. These parameters were used to calculate the explained common variance (ECV) index and the item explained common variance (IECV) index (Rodriguez et al., 2016). Large values of ECV (i.e., > 0.85), suggest that the set of items can be reasonably considered unidimensional.

The ECV for the overall job crafting scale dimension is 0.37 and 0.39 for the general and daily versions, respectively. This means that the overall job crafting dimension accounts for 37% and 39% of the variance accounted for by the bifactor model as a whole. The seeking resources, seeking challenges, and reducing demands dimensions accounted for 12%, 23%, and 28% of the variance, respectively, for the general version and 20%, 10%, and 30% for the daily version. The $IECV_G$ was also computed for each item on the general dimension of the bifactor model (see Tables E4 and E5 of ESM 1) with almost all values below 0.85.

Discussion

In this brief report, we used MIRT to evaluate the factor structure of the Petrou et al. (2012) job crafting scale. The study contributes to research by clarifying that an overall factor of job crafting does not exist in either of the two-scale versions. Based on our MIRT analyses the correlated factors model provided an adequate global fit to the data. As such, we suggest that the Greek version of the PJCS is best conceptualized as being defined by three independent, yet related dimensions.

From a theoretical perspective, our empirical results do not support the existence of a measurable "overall job crafting" construct at the general or daily levels. This finding is in line with recent studies (i.e., Bakker et al., 2018) using Classical Test Theory (CTT) techniques on different constructs for the assessment of job crafting and recent meta-analyses showing that different dimensions of job crafting have different antecedents and outcomes (Lichtenthaler & Fischbach, 2019).

Regarding limitations the study used self-reported measures and as such, self-report bias and common method variance may have affected the results. Future research on the structure of job crafting could benefit of the use of reports made by colleagues or supervisors. Furthermore, our sample was composed of Greek high educated employees; generalizability to employees with less formal education remains an open question. It is plausible that less-educated employees may participate in job crafting behaviors in different levels and forms compared to more educated employees (e.g., engage in crafting behaviors that are protective of their status and function in their work).

Moreover, we implicitly theorized that our analyses are based on a compensatory MIRT model (Reckase, 2009), meaning that a low score in one dimension can be compensated by a high score in another dimension. Future research could use non-compensatory MIRT models and examine how overall scores may relate to external variables included in job crafting's nomological network.

Is there any merit to developing direct measures of overall job crafting? Job crafting is an important proactive approach to job redesign that facilitates the work-related well-being of employees. Theoretically, we could expect that different dimensions of job crafting to interact to determine employee behaviors (in line with our compensatory MIRT model conceptualization) and recent empirical findings confirm this (see Petrou & Xanthopoulou, 2020) for interactive effects of job crafting dimensions). More research is clearly needed on whether and which multiple dimensions of job crafting can be represented as an aggregate score.

Conclusion

The results of the present study, caution against aggregated scores across the dimensions of job crafting for both the general and daily version of the PJCS. The PJCS

conceptualizes job crafting as a multidimensional proactive behavior, and as such, the use of alternative operationalization of job crafting (such as an overall aggregated scale score) may lead empirical research to very different conclusions regarding antecedents and outcomes.

Electronic Supplementary Materials

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1015-5759/a000638

ESM 1. General all models

References

Akkermans, J., & Tims, M. (2017). Crafting your career: How career competencies relate to career success via job crafting. *Applied Psychology: An International Review, 66*(1), 168–195. https://doi.org/10.1111/apps.12082

Anderson, D., Kahn, J. D., & Tindal, G. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Applied Measurement in Education, 30*(3), 163–177. https://doi.org/10.1080/08957347.2017.1316277

Bakker, A. B., Ficapal-Cusí, P., Torrent-Sellens, J., Boada-Grau, J., & Hontangas-Beltrán, P. M. (2018). The Spanish version of the Job Crafting Scale. *Psicothema, 30*(1), 136–142. https://doi.org/10.7334/psicothema2016.293

Cai, L., Thissen, D., & du Toit, S. (2017). *IRTPRO for Windows*. Scientific Software International

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Press.

Lichtenthaler, P. W., & Fischbach, A. (2019). A meta-analysis on promotion- and prevention-focused job crafting. *European Journal of Work and Organizational Psychology, 28*(1), 30–50. https://doi.org/10.1080/1359432X.2018.1527767

Petrou, P., Demerouti, E., Peeters, M. C., Schaufeli, W. B., & Hetland, J. (2012). Crafting a job on a daily basis: Contextual correlates and the link to work engagement. *Journal of Organizational Behavior, 33*(8), 1120–1141. https://doi.org/10.1002/job.1783

Petrou, P., & Xanthopoulou, D. (2020). Interactive effects of approach and avoidance job crafting in explaining weekly variations in work performance and employability. *Applied Psychology: An International Review*. https://doi.org/10.1111/apps.12277

Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045

Rudolph, C. W., Katz, I. M., Lavigne, K. N., & Zacher, H. (2017). Job crafting: A meta-analysis of relationships with individual differences, job characteristics, and work outcomes. *Journal of Vocational Behavior, 102*, 112–138. https://doi.org/10.1016/j.jvb.2017.05.008

Tims, M., & Bakker, A. B. (2010). Job crafting: Towards a new model of individual job redesign. *South African Journal of Industrial Psychology, 36*, 1–9.

Tims, M., Bakker, A. B., & Derks, D. (2012). Development and validation of the Job Crafting Scale. *Journal of Vocational Behavior, 80*(1), 173–186. https://doi.org/10.1177/20018726712453471

History
Received February 1, 2020
Revision received November 23, 2020
Accepted December 10, 2020
Published online March 2, 2021
EJPA Section / Category I/O Psychology

Open Data
The data that were analyzed for this paper and the supplemental files are available via the Open Science Framework: https://osf.io/gka4s/?view_only=f8fe6616d1de41ce9792bde2e0a8f554

ORCID
Leonidas A. Zampetakis
https://orcid.org/0000-0002-9568-6302

Leonidas A. Zampetakis
Applied Psychology Laboratory
Department of Psychology
University of Crete
Gallos University Campus
Rethymno
Crete 74100
Greece
zampetakis.leonidas@uoc.gr

Multistudy Report

It's All About Power

Validation of Trait and State Versions of the German Personal Sense of Power Scale

Robert Körner[1,2], Timo Heydasch[3], and Astrid Schütz[2]

[1]Department of Psychology, Martin-Luther-University of Halle-Wittenberg, Germany
[2]Department of Psychology, Otto-Friedrich-University of Bamberg, Germany
[3]Department of Psychology, Distance University of Hagen, Germany

Abstract: The present research was aimed at providing a German version of the Personal Sense of Power Scale (GPSPS; Anderson et al., 2012) and testing its psychometric properties. A personal sense of power describes the perception of one's ability to influence others. Probably every human relationship can be characterized by differences in power, which means that the measurement of experienced power is highly relevant. The availability of appropriate measures in different languages will help improve research and cross-cultural comparisons. Five studies were conducted. Internal consistency was high across all studies. Stability across 6 and 12 weeks was also high. A good fit was observed for a 6-item unidimensional version. Correlations with a variety of psychological and sociodemographic variables were in the expected directions, supporting nomological and criterion validity (Study 1). Measurement invariance across gender was demonstrated. In support of construct validity, a clinical sample scored significantly lower than others. Finally, two studies showed the sensitivity of a state version of the scale. We encourage researchers to use this scale as a reliable and valid instrument for assessing trait and state power.

Keywords: power, personal sense of power, state power, influence, status

"The fundamental concept in social science is Power, in the same sense in which Energy is the fundamental concept in physics" (Russell, 1938, p. 10). Russell's statement can be found in various articles on power and status and illustrates the importance of power in psychological research and everyday life. In recent decades, several intriguing theories have emerged (e.g., Keltner et al., 2003; Magee & Smith, 2013), and various findings have been published. Power has overcome the stigma of being connected to only negative outcomes (e.g., corruption, self-serving behavior, and egocentric biases). Instead, power can be seen as an intensifier of goal-related approach motivation (Guinote, 2017). Accordingly, Guinote's (2017) review shows that power energizes thought, speech, and action, increases prioritization and authenticity, but also leads to stereotyping and objectification. Thus, on the basis of predispositions and situational circumstances, power apparently intensifies people's behavioral tendencies in either antisocial or prosocial ways.

In social psychology, power is often described as a type of resource control that can modify others' states (Keltner et al., 2003). Yet, power can be independent of sociostructural factors: Anderson et al. (2012) thus defined a subjective sense of power as a "psychological state – a perception of one's capacity to influence others" (p. 314). For example, an employee might make decisions in a negotiation despite lacking a formal position and responsibility. Thus, the employee might experience a high personal sense of power even without the formal position. But how can the experience of power be measured? We aimed to provide and validate a German version of the only established measure of generalized power: The Personal Sense of Power Scale (PSPS; Anderson et al., 2012).

The 8-item unidimensional PSPS captures individuals' beliefs about their influence over others and their decision-making ability within social relationships. Using nine different samples, Anderson et al. (2012) reported high internal consistency for the scale and showed a distinct but moderately related personal sense of power between different relationship types (e.g., friend relationship, parent relationship). Further, they demonstrated the existence of a personal sense of power for different abstraction levels: short-term and long-term dyadic relationships, groups, and a generalized form.

The PSPS has become very popular in a very short amount of time. The scale has been in use since the early 2000s (e.g., Anderson & Galinsky, 2006), and its theory

and development were presented in 2012 (Anderson et al., 2012). Anderson et al. (2012) presented instructions for the PSPS for different relationship types (e.g., date-, supervisor-, friend-relationships). As of October 2020, the original publication has been cited more than 600 times (Google Scholar). The scale has been translated into several languages such as Chinese (e.g., Wang, 2015), Dutch (e.g., Van Kleef et al., 2015), Hebrew (Uziel & Hefetz, 2014), and Polish (e.g., Kocur & Mandal, 2018), and acceptable internal consistencies have been reported for these translations. Researchers have also used the measure in Germany (e.g., Weineck et al., 2019).[1] Yet, to the best of our knowledge, the PSPS has not been validated in any language other than English. In the present study, we aimed to identify the psychometric properties of the German version of the Personal Sense of Power Scale (GPSPS), test the scale in distinct samples, extend predictions regarding its validity, and for the first time, test the unidimensionality of the scale by applying confirmatory factor analyses and examine the measurement invariance of the scale across sex.

Another important aspect of a personal sense of power is that it has been used for manipulation checks and as a predictor and an outcome variable. As the PSPS is usually conceptualized as a trait measure, researchers have sometimes found no effect of an experimental power manipulation on this scale (e.g., Deuter et al., 2016). Therefore, in the current study, we also aimed to test and establish instructions for a state version of the GPSPS to measure situational fluctuations in personal power.

Overview of Studies

We conducted five studies to provide an in-depth examination of the GPSPS's psychometric properties. Studies 1–3 were designed to test the unidimensionality of the trait version of the GPSPS with confirmatory factor analysis (CFA). In Study 1, we tested the scale's internal consistency and stability and assessed a variety of psychological and sociodemographic constructs for nomological and criterion validity. Further, we tested for measurement invariance across gender. In Study 2, we used a community sample and measured personal sense of power in the context of romantic relationships to further test for internal consistency and unidimensionality. A clinical sample was used in Study 3 to make a comparison between groups (i.e., clinical and nonclinical groups). Finally, in Studies 4 and 5, we tested a state version of the GPSPS.

Study 1

The first study was aimed at examining the reliability and unidimensionality of the GPSPS and at providing detailed information about nomological and criterion validity. The GPSPS was based on a translation/back-translation procedure. The scale was used as a trait measure reflecting a generalized sense of power: "In my relationships with others. . .."

To test the nomological and criterion validity of the PSPS, we relied on the variables and measures used by Anderson et al. (2012) but also added several new measures (e.g., facets of narcissism, construal style). On the basis of the literature, we expected positive associations between a personal sense of power and extraversion, conscientiousness, openness (Anderson & Cowan, 2014), internal locus of control (Anderson et al., 2012), dominance (Anderson & Cowan, 2014; Dunbar & Burgoon, 2005), narcissism (Brunell et al., 2008), self-esteem (Körner et al., 2019; Wojciszke & Struzynska-Kujalowicz, 2007), and behavioral activation (Keltner et al., 2003). A personal sense of power was not expected to be associated with agreeableness, and negative associations were expected with neuroticism (Anderson & Cowan, 2014) and behavioral inhibition (Keltner et al., 2003).

Further, to extend Anderson et al.'s (2012) nomological network of personal power on the basis of two major theories in the field of power, we made the following predictions: Positive emotions will be positively correlated, and negative emotions will be negatively correlated with a personal sense of power (approach-inhibition theory of power; Keltner et al., 2003). With respect to the social distance theory of power (Magee & Smith, 2013), positive associations were expected for abstract construal style and social distance. Finally, as pride is the emotion most closely linked to social rank (Cheng et al., 2010), we also expected a positive association between power and pride. Yet, authentic pride should show a stronger association with a personal sense of power than hubristic pride because associations between personality variables and authentic pride are similar to a personal sense of power.

Moreover, we made some predictions regarding criterion validity. The original publication did not test for associations between objective criteria and personal sense of power. As the experience of power may be independent of sociostructural aspects but usually does show a moderate relation, we expected a positive but small correlation

[1] Weineck et al. (2019) used only six items (Items 1–6) from the original scale but these were different from the items that we had identified as being psychometrically adequate (Items 1, 2, 4, 5, 6, and 7). They reported a Cronbach's α of .82, which is slightly below the mean Cronbach's α reported in the present studies (M_α = .85).

between subjective power and socioeconomic status (Anderson et al., 2012), managerial responsibility (Boeker, 1992), and number of employees. Given that status is associated with increased body height (Stulp et al., 2012), and powerful people overestimate their body height (Duguid & Goncalo, 2012), we also expected a positive association between body height and sense of power.

Method

Participants and Procedure

Participants were recruited online at a distance-learning university to collect data from a more heterogeneous sample with respect to age and professional background. Participants were offered course credit for completing the questionnaires. They lived all over Germany. We examined the stability of the GPSPS across three points of measurement. The questionnaire used at the first time point (t1) consisted of the GPSPS and several measures that were included to establish validity. Participants generated an individual code so that retest results could be matched. After data preparation (see Results section), the sample comprised 573 participants (80% women, 19% men, 1% diverse; M_{age} = 32.12, SD_{age} = 10.16, range: 18–75 years). After 6 weeks (t2), 266 individuals completed the GPSPS for a second time (80% women, 18% men, 1% diverse; M_{age} = 33.46, SD_{age} = 10.83, range: 18–75 years). Finally, 185 participants completed the scale for a third time after 12 weeks (t3; 79% women, 18% men, 1% diverse, M_{age} = 33.75, SD_{age} = 11.02, range: 18–75 years). We also tested for whether there was a pattern in the missing data across measurement points. Little's missing completely at random (MCAR) tests were not significant for the comparisons of the GPSPS scores, $\chi^2(2) = 2.627$, $p = .269$ (t1 with t2), $\chi^2(2 = 1.676$, $p = .432$ (t1 with t3), and $\chi^2(1) = 0.376$, $p = .540$ (t2 with t3). This supported the null hypothesis that the data were missing completely at random.

Study 1 was preregistered (http://aspredicted.org/blind.php?x=429eg5). Codes and data for all studies are available at https://osf.io/jf9dz. Correlational analyses and group comparisons were done with SPSS 25. Factor analyses were computed with Mplus 7 (Muthén & Muthén, 1998–2012). In the CFAs for all studies, all error terms were uncorrelated. RStudio 1.2.5019 was used to calculate McDonald's ω. For all studies, we report how we determined our sample size, all data exclusions, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact p values, effect sizes, and 95% confidence or credible intervals.

Measures

The PSPS (Anderson et al., 2012) comprises eight items (e.g., "My ideas and opinions are often ignored") rated on a 7-point scale ranging from 1 (= *strongly disagree*) to 7 (= *strongly agree*). We used a translation/back-translation procedure to create the German version according to the Guidelines for Translating and Adapting Tests by the International Test Commission (2017). First, two experts in psychological power research translated the items into German. A bilingual native English speaker back-translated the items. There was high congruence in wording. Minor discrepancies occurred and were resolved in a discussion. The items and response format can be found in Table 1. Cronbach's α coefficients are presented in Table 2 for all scales.

Various trait measures were used to assess nomological validity. The habitual experience of positive and negative emotions was measured with the Positive and Negative Affect Schedule (PANAS; German version: Krohne et al., 1996). Participants were asked to use a 5-point rating scale ranging from 1 (= *not at all*) to 5 (= *extremely*) to rate the extent to which they generally experienced 20 emotions. Half of the items addressed positive affect (e.g., excited) and the other half negative affect (e.g., ashamed).

The 7-item Authentic and Hubristic Pride Scale (Tracy & Robins, 2007) covers two facets of pride: Authentic pride refers to confidence and success (e.g., "I feel I am achieving"), whereas hubristic pride refers to arrogance and conceitedness (e.g., "I am smug"). The scale was administered with a 5-point rating scale ranging from 1 (= *not at all*) to 5 (= *extremely strong*).

The Rosenberg Self-Esteem Scale (German version: von Collani & Herzberg, 2003) measures trait self-esteem with 10 items (e.g., "I certainly feel useless at times"). Answers were given on a rating scale ranging from 1 (= *strongly disagree*) to 5 (= *strongly agree*).

Narcissism was measured with the short-form of the Narcissistic Personality Inventory (NPI-15; German version: Schütz et al., 2004). The 15-item scale addresses subclinical grandiose narcissism as a personality trait. The items have a dichotomous forced-choice format. One statement from each pair represents narcissism (e.g., "Everyone likes to listen to me"). Further, we used the short form of the Narcissistic Admiration and Rivalry Questionnaire (NARQ; Back et al., 2013). People who want to be admired by others for the purpose of self-exaltation score high on Admiration (e.g., "I deserve to be considered a great person"). Rivalry addresses asserting oneself against others to protect oneself (e.g., "I want my competitors to fail"). Each facet consists of three items. Answers were given

Table 1. Descriptive statistics, corrected item-total correlations (r_{it}), and loadings of the GPSPS items in Study 1

Item	*M*	*SD*	r_{it}	Loading
1. Ich bekomme Menschen dazu, mir zuzuhören.*	5.55	1.09	.59	.68
[I can get him/her/them to listen to what I say.]				
2. Meine Wünsche haben nicht viel Gewicht.[R]*	5.06	1.39	.60	.69
[My wishes do not carry much weight.]				
3. Ich kann Menschen dazu bringen, zu tun, was ich will.	4.78	1.26	–	–
[I can get him/her/them to do what I want.]				
4. Auch wenn ich meine Ansichten ausspreche, haben diese wenig Einfluss.[R]*	5.24	1.22	.73	.84
[Even if I voice them, my views have little sway.]				
5. Ich habe viel Macht.*	3.57	1.37	.53	.59
[I think I have a great deal of power.]				
6. Meine Ideen und Meinungen werden oft ignoriert.[R]*	5.27	1.29	.71	.83
[My ideas and opinions are often ignored.]				
7. Selbst wenn ich es versuche, kann ich mich nicht durchsetzen.[R]*	5.54	1.26	.69	.81
[Even when I try, I am not able to get my way.]				
8. Wenn ich will, dann treffe ich die Entscheidungen.	5.25	1.27	–	–
[If I want to, I get to make the decisions.]				

Note. *Final items; [R]Inverse items. Response format: 1 = *strongly disagree* (stimme gar nicht zu), 2 = *largely disagree* (stimme kaum zu), 3 = *somewhat disagree* (stimme eher nicht zu), 4 = *neither* (weder noch), 5 = *somewhat agree* (stimme eher zu), 6 = *largely agree* (stimme weitgehend zu), 7 = *strongly agree* (stimme völlig zu). The original English items are shown in brackets below each German item and are reprinted here from Anderson et al. (2012) with the permission of the authors.

Table 2. Nomological validity of the GPSPS: Descriptive statistics for the dependent measures and zero-order correlations with personal sense of power

Dependent measure	Cronbach's α	*N*	*M*	*SD*	Range	Expected correlation	Observed correlation
Positive emotions	.85	569	3.44	0.63	1–5	+	.44***
Negative emotions	.87	569	1.98	0.67	1–5	–	–.38***
Authentic pride	.89	569	3.47	0.75	1–5	+	.52***
Hubristic pride	.85	569	1.81	0.66	1–5	+	.12**
Self-esteem	.90	565	3.15	0.59	1–4	+	.52***
Narcissism (NPI)	.78[a]	567	4.90	3.21	0–15	+	.49***
Narcissism (NARQ)	.79	569	2.62	0.93	1–6	+	.23***
Admiration	.80	569	2.93	1.17	1–6	No pre	.34***
Rivalry	.66	569	2.31	0.98	1–6	No pre	.04
Dominance	.67	568	5.11	1.06	1–8	+	.60***
Openness	.78	565	3.80	0.77	1–5	+	.07*
Conscientiousness	.75	565	3.97	0.61	1–5	+	.25***
Extraversion	.76	565	3.24	0.71	1–5	+	.39***
Agreeableness	.73	565	3.91	0.67	1–5	0	–.02
Neuroticism	.84	565	2.57	0.85	1–5	–	–.54***
Internal locus of control	.76[a]	566	11.58	4.27	0–23	+	.25***
Behavioral activation	.75	586	3.08	0.36	1–4	+	.30***
BAS Drive	.69	586	3.07	0.49	1–4	+	.28***
BAS Fun Seeking	.59	569	2.91	0.50	1–4	+	.11**
BAS Reward Responsiveness	.60	569	3.23	0.44	1–4	+	.28***
Behavioral inhibition	.83	569	2.97	0.56	1–4	–	–.36***
Abstract construal style	.87[a]	565	15.75	5.54	0–25	+	.17***
Social distance	–	565	3.93	1.55	1–7	+	.14**

Note. [a]Values were calculated with the Kuder-Richardson Formula 20. BAS = behavioral activation system; no pre = no prediction was made for this variable in the preregistration. *$p < .05$; **$p < .01$; ***$p < .001$ (all one-tailed).

on a 6-point scale ranging from 1 (= *strongly disagree*) to 6 (= *strongly agree*).

To measure dominance, we used adjectives from the Revised Interpersonal Adjective Scales (IAS-R; Wiggins et al., 1988). We relied on the findings by Lorr and Strack (1990), who identified seven adjectives (e.g., "assertive") that were the best markers for the dominance-submission dimension. Answers were given on an 8-point rating scale ranging from 1 (= *extremely inaccurate*) to 8 (= *extremely accurate*) regarding how the person feels in general.

The NEO-FFI-30 (Körner et al., 2008) is a German short form of the NEO Five-Factor Inventory and measures the Big Five with six items each. Answers were given on a 5-point rating scale ranging from 1 (= *strongly disagree*) to 5 (= *strongly agree*).

Locus of control was measured with the Internal-External Control scale (German version: Rost-Schaude et al., 2014). The 28 items (five filler items) have a dichotomous forced-choice format. One statement represents internal and the other external locus of control (e.g., "Unfortunately, a person's values often go unrecognized, no matter how hard he tries").

The BIS/BAS Scale (German version: Strobel et al., 2001) consists of 24 items with two superior factors: behavioral activation (BAS) and behavioral inhibition (BIS; e.g., "Criticism or scolding hurts me quite a bit"). The BAS factor can be divided into three components: Fun Seeking (e.g., "I am always willing to try something new if I think it will be fun"), Drive (e.g., "I go out of my way to get things I want"), and Reward Responsiveness (e.g., "It would excite me to win a contest"). Answers were given on a 5-point rating scale ranging from 1 (= *strongly disagree*) to 4 (= *strongly agree*).

The Behavior Identification Form (Vallacher & Wegner, 1989) measures construal style with 25 items. Participants were informed that behaviors can be identified in different ways. Then they had to choose one of two alternatives for certain behavior (e.g., "making a list: (a) getting organized versus (b) writing things down" representing (a) a high-level identity or (b) a low-level identity).

Social distance was measured with the single-item measure Inclusion of Other in the Self Scale (Aron et al., 1992). Participants were instructed to circle the diagram that best described their interpersonal relationships. Each diagram consisted of two circles labeled "self" and "other." Answers were given on a pictorial 7-point rating scale ranging from 1 (= *circles for self and other do not overlap*) to 7 (= *circles for self and other almost completely overlap*).

Several sociodemographic characteristics were measured: age, gender, body height (in cm), managerial responsibility, and a number of employees. Further, profession, net income, and educational and vocational qualifications were measured to assess sociodemographic status (for the procedure, see Lampert et al., 2013).

Results

Data Preparation

At t1, the questionnaire was completed by 607 participants. To ensure the quality of the data and the validity of the protocol (see Johnson, 2005), we conducted different data-cleaning steps in accordance with our preregistration. First, we excluded 11 participants with an average answer time below 2 s per item. Next, the individual reliability coefficient (IRC; Jackson, 1976) of the remaining 596 cases was computed using scales with more than one item, whereby the scales were adjusted according to the different rules for computing the scales (e.g., mean vs. sum; item coding zero to one vs. one to five). Five participants were excluded because they had an IRC below zero. The remaining 591 cases were examined to identify patterns of vertical answering, that is, they almost always provided the same score across items (e.g., agreeing strongly even when the items were inverted or referred to different matters). The percentage of consecutive identical answers (PCIA; Heydasch, 2014) was calculated (the number of consecutive identical answers on a rating scale divided by the number of items using that rating scale multiplied by 100). To obtain an overview, we averaged the PCIAs of all rating scales and excluded three participants who had nearly always chosen the same option (PCIA > 90%). Finally, as planned in the preregistration, 15 cases in which individuals participated repeatedly with an identical code were deleted. In total, 573 valid cases remained in the sample and were used in the statistical analyses.

Factorial Validity and Item Characteristics

As assessed with the Kolmogorov-Smirnov test ($ps < .001$) and the Shapiro-Wilk test ($ps < .001$), the items and the sum score for the GPSPS were not normally distributed. Thus, we used the weighted least squares estimator (WLSMV) for the CFA (DiStefano & Hess, 2005). The expected unidimensional factor solution showed fit indices that were not satisfactory, $\chi^2(20) = 240.982$, $p < .001$; RMSEA = .139, 90% CI [.123, .155], $p < .001$; CFI = .955; TLI = .937. We then examined the modification indices and identified two items that were responsible for the poor fit (Items 3 and 8). The items were both about "wanting something" and thus differed from the rest of the items. The resulting 6-item factor solution showed good fit, $\chi^2(9) = 22.454$, $p < .001$; RMSEA = .051, 90% CI [.025, .078], $p = .430$; CFI = .997; TLI = .995. All loadings were significant ($ps < .001$). In the following, we used the 6-item version. Table 1 presents the means, standard deviations, and corrected item-total correlations for the items.

Reliability

The split-half reliability was acceptable at .74 (Items 1, 2, and 4 correlated with Items 5, 6, and 7). Cronbach's α for the GPSPS was good at .85 (.86 at t2 and t3). McDonald's ω was computed by using the robust maximum-likelihood estimator (MBESS package in R; Kelley, 2018), and there was also good internal consistency at .85 (.87 at t2 and t3).

Stability

We found high test-retest correlations for the 6-week, $r_{t1t2}(264) = .74$, $p < .001$, and 12-week intervals, $r_{t1t3}(183) = .72$, $p < .001$.

Nomological Validity

All associations between the GPSPS and the psychological scales were in the expected directions (see Table 2). Interestingly, the correlation with authentic pride was much higher than with hubristic pride ($z = 8.00$, $p < .001$). High positive correlations were found for the GPSPS with self-esteem, $r(563) = .52$, $p < .001$, and dominance, $r(566) = .60$, $p < .001$. With respect to narcissism, there was a positive association with admiration, $r(567) = .34$, $p < .001$, but no association with rivalry, $r(567) = .04$, *ns*. The strongest correlation with the Big Five was for neuroticism, $r(563) = -.54$, $p < .001$. The association with openness was positive as expected but almost zero, $r(563) = .07$, $p < .05$. For the facets of behavioral activation, the GPSPS showed higher correlations with drive ($z = 3.31$, $p < .001$) and reward responsiveness ($z = 3.67$, $p < .001$) than with fun seeking. There were also small but significant positive relations with abstract construal style, $r(563) = .17$, $p < .001$, and social distance, $r(563) = .14$, $p < .01$.

Criterion Validity

The GPSPS's associations with socioeconomic status and managerial responsibility were in the expected directions (see Table 3). The GPSPS's correlation with number of employees was unexpectedly close to zero, $r(566) = -.03$, $p = .235$. However, an inspection of the *z*-transformed data for the employee variable showed an outlier ($z = 10.06$ with 600 employees). This person was excluded, and the GPSPS's association with the number of employees became slightly larger, $r(565) = .08$, $p = .036$. When excluding participants who supervised more than 50 employees (cut-off for small companies) or more than 10 employees (cut-off for microenterprises), the association increased, $r(560) = .11$, $p = .004$, $r(533) = .16$, $p < .001$, respectively. Unexpectedly, there was no clear relation between the GPSPS and body height (see Table 3). The correlation between body height and sense of power was for men, $r(107) = .08$, and for women, $r(454) = .01$.

Table 3. Zero-order correlations between the GPSPS and sociodemographic characteristics

Dependent measure	Expected correlation	Observed correlation
Age	No pre	.10*
Gender[a]	No pre	−.07
Body height	+	.04
Socioeconomic status	+	.18***
Managerial responsibility	+	.20***
Number of employees	+	−.03

Note. No pre = no prediction was made for this variable in the preregistration. [a]Male = 1, Female = 2. *$p < .05$; ***$p < .001$ (all one-tailed).

Measurement Invariance

We tested for measurement invariance across gender (only male and female). Using multigroup CFA, we found strict invariance for the GPSPS (see Table 4) with respect to the invariance criterion by Cheung and Rensvold (2002; $\Delta CFI \leq .01$).

Discussion

The results largely supported the preregistered expectations. The GPSPS showed a unidimensional structure and good fit with six items. Two items were excluded. The modification indices suggested that adding covariances between Items 3, 8, and the other six items would improve the fit of the model. As correlated error terms violated the assumption of local model fit in a unidimensional model, the best approach was to remove these two items from the final scale. Further, Item 8 also showed the lowest corrected item-total correlation as well as the lowest loading in the CFA (see the Online Supplementary Material at https://osf.io/2tqwc/). Cronbach's α barely changed when Items 3 and 8 were excluded. With respect to the content, the two items seemed to have something in common (they are about "wanting something") – an aspect that is not present in the other items. This suggests that these items may represent a different latent variable. The final GPSPS items showed high corrected item-total correlations. Internal consistency was satisfactory and similar to the values found for the original scale. The trait version showed high stability.

The construct was correlated with other variables in the expected directions. The strongest association was with dominance, which is a closely related construct with respect to social hierarchy. Also, its association with authentic pride, which is also closely related to power (Cheng et al., 2010), was expected. Self-esteem and narcissism also showed strong positive correlations with the personal sense of power, which suggests that this sense is linked to overall positive self-evaluations. Neuroticism showed the strongest negative association with a personal sense of power, which suggests that emotional stability could lead to or might be a

Table 4. Test of measurement invariance for gender (Male/Female) in Study 1 (t1)

Fit indices	Configural	Metric	Scalar	Strict (factor variances)	Strict (residual error variances)
χ^2	35.448	43.902	55.813	56.729	74.733
RMSEA	.059	.057	.057	.056	.062
90% CI	[.029, .087]	[.030, .082]	[.034, .080]	[.033, .078]	[.042, .081]
CFI	.987	.984	.980	.980	.971
TLI	.978	.980	.979	.980	.976
AIC	9,911.725	9,910.178	9,910.089	9,909.005	9,915.009
BIC	10,067.850	10,044.210	10,018.510	10,013.089	9,993.072

Note. RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; AIC = Akaike Information Criterion; BIC = Bayes Information Criterion.

consequence of personal power. Of course, third variables such as depression or anxiety may be the basis for this association. This finding dovetails with the associations found with positive and negative emotions. Further, the expected correlations (emotions, behavioral activation, and inhibition) with respect to the approach/inhibition theory of power (Keltner et al., 2003) were high. Interestingly, however, the correlations with construal style and social distance were only small to medium in size. Overall, this may suggest that the GPSPS has a better match with the nomological net as proposed by the approach/inhibition theory than with the associations suggested by the social distance theory of power (Magee & Smith, 2013). Moreover, the present patterns and sizes of the correlation coefficients were largely comparable to the findings from the original scale (Anderson et al., 2012). Only the association with neuroticism was much stronger in the present study than it was in the original study, and the association with openness was much weaker. When potential cross-cultural differences are taken into account, this may suggest that emotional stability is more decisive for decision-making ability in Germany than in the US. But another way to explain these differences might be that the Big Five items have slightly different meanings in English and German (Hofstee et al., 1997).

Criterion validity was supported as the GPSPS showed small but positive associations with aspects of sociostructural power. However, the association between GPSPS and body height was not as expected. Apparently, physical features do not necessarily correspond to a personal sense of power. Despite a great deal of literature suggesting that body height is positively associated with power and status (e.g., Stulp et al., 2012), there are studies that have shown no association (e.g., between body height and earnings in Germany; Heineck, 2005). Moreover, the overrepresentation of women in the sample may have prevented an association between sense of power and body height from being found. In fact, the association between sense of power and height is somewhat stronger for men than for women. Finally, because strict measurement invariance was established, the personal sense of power was measured in the same way for both men and women.

Study 2

In Study 2, we cross-validated the unidimensional factor structure with six items in a second sample and assessed internal consistency. We used the GPSPS in the context of romantic relationships because the sense of power is considered to pertain to various types of contexts and relationships (Anderson et al., 2012). We thus aimed to increase the applicability of the scale across contexts. The instruction read: "In the relationship with my partner…."

Method

Undergraduates of a university course recruited participants via the snowball principle. Participants mostly were from southern Germany. Participants could participate online or offline. There was no incentive for participation. Overall, 435 participants took part (54% women, 46% men; M_{age} = 30.39, SD_{age} = 12.84, 14 to 73). All participants were in a romantic relationship (23.9% married, 3.4% engaged, 72.6% dating). The average relationship duration was 8 years (SD = 10.39, range: 1 month to 52 years).

Results and Discussion

As in Study 1, the 6-item GPSPS showed an acceptable fit, $\chi^2(9) = 55.988$, $p < .001$; RMSEA = .110, 90% CI [.083, .138], $p < .001$; CFI = .976; TLI = .961. Reliability was acceptable when computed as Cronbach's α (α = .78) or McDonald's ω (ω = .80). Further, the model fit the data much better than the 8-item version, $\chi^2(20) = 463.656$, $p < .001$; RMSEA = .226, 90% CI [.208, .244], $p < .001$; CFI = .806; TLI = .728. Overall, the CFA supported the one-factor solution in a second independent community

sample with better gender representation. Yet, the RMSEA was slightly above the traditional cut-off values for acceptable fit. This may have occurred because the violation of multivariate normality was largest in this sample (particularly with a kurtosis value > 3 for Item 1) and the degrees of freedom were low (Hammervold, 1998; Kenny et al., 2015). Because the CFI and TLI showed acceptable values and the RMSEA was acceptable in Studies 1 and 3, we concluded that the 6-item solution was preferable.

Study 3

In this study, we examined the factorial validity of the GPSPS in a clinical sample. Moreover, we tested for construct validity: As individuals with mental disorders show impairments in their decision-making ability and their volitional control (Goschke, 2014), it seems plausible that they would experience a lower personal sense of power in their general relationships than others. Many patients experience stigma or discrimination due to their mental illness and consequently report lower personal power (Lysaker et al., 2008; Mashiach-Eizenberg et al., 2013). In addition, other proxies of personal power, or the lack of it, such as behavioral inhibition, a prevention focus (Keltner et al., 2003), or neuroticism as found in Study 1, are associated with an increased likelihood of developing a mental disorder (Clauss & Blackford, 2012; Eddington et al., 2009; Lahey, 2009). To the best of our knowledge, such a test of extreme group validity has not been previously reported for the scale, but as elaborated above, it makes conceptual sense for impairment to be associated with a lack of experienced power. The GPSPS was used as a trait measure to measure a generalized sense of power: "In my relationships with others…."

Method

Participants were recruited online via 10 communities and fora concerning mental disorders, depression, and self-help. As an incentive, participants could be entered into a drawing for Amazon vouchers. The questionnaire contained items on demography and psychotherapeutic indications and the trait GPSPS. A total of 187 individuals participated; two were excluded due to vertical answer patterns; two responded too quickly (see Leiner, 2013). The final sample comprised 183 participants (77.6% women, 16.4% men, 1.6% diverse; M_{age} = 37.31, SD_{age} = 13.66, range: 16–83 years). Eighty-nine participants (48.6%) were currently in psychotherapeutic treatment; 157 (85.8%) reported at least one diagnosed mental disorder; 87 (47.5%) reported more than one diagnosed mental disorder. The following mental disorders were named: major depression (77.7%), anxiety disorders (33.8%), trauma- and stress-related disorders (24.2%), and borderline personality disorder (19.8%). This study was not preregistered as we were not able to estimate a priori how many participants would end up participating in this study.

Results and Discussion

First, missing values were replaced with the expectation-maximization method. Little's MCAR test was not significant, $\chi^2(28) = 24.393$, $p = .661$, which suggested that the data were missing completely at random. A total of six missing values were replaced. Internal consistency was high ($\alpha = .88$, $\omega = .88$). Then, a CFA was computed. The expected unidimensional factor solution fit the data well, $\chi^2(9) = 21.909$, $p < .01$; RMSEA = .089, 90% CI [.042, .136], $p = .081$; CFI = .994; TLI = .990. Finally, we compared the mean of the GPSPS in this sample with the mean of the GPSPS in the sample from Study 1 (t1). An ANCOVA controlling for age and gender showed the expected main effect, $F(1, 736) = 155.207$, $p < .001$, $\eta_p^2 = .17$. The participants in the Study 1 sample reported a significantly higher personal sense of power ($M = 5.04$, $SD = 0.97$) than the clinical sample participants ($M = 3.91$, $SD = 1.28$). When we excluded participants from Sample 3 who had not indicated a diagnosed mental disorder, the effect size increased, $F(1, 711) = 154.886$, $p < .001$, $\eta_p^2 = .18$ (Sample 3: $M = 3.86$, $SD = 1.26$).

To sum up, high reliability was found in a third and clinical sample, and the unidimensional structure and fit of the GPSPS were supported. Moreover, participants who reported diagnosed mental disorders had a lower personal sense of power than others, which provides initial support for the measure's construct validity. Yet, we had not asked for mental disorders in Study 1, which allows for the possibility that some of the Sample 1 participants might also suffer from a disorder. Furthermore, hospitalized patients with major mental health issues were not included in our clinical sample. Consequently, the differences between the clinical and non-clinical populations may in fact be even larger.

Study 4

The aim of Study 4 was to test a state version of the GPSPS. So far, the instructions for the PSPS have been trait-oriented. By contrast, in experimental designs concerning power, researchers have typically used individual items to measure experienced power. Yet, a validated scale

to measure the state experience of power is helpful as it provides the opportunity for parallel measurement of state and trait power and increasing measurement accuracy. We used a simple method to transform the GPSPS into a state version: We used instructions that are often used for state measures. To test the validity of the instructions and the state GPSPS, we used an often-employed intervention in power research: autobiographical recall (e.g., Galinsky et al., 2003). Participants were assigned to a high- or a low-power group only because we were interested in the sensitivity of the scale. The instructions for state sense of power read: "Please tick the option that applies most to you at the moment."

Method

As stated in the preregistration (https://aspredicted.org/blind.php?x=8n4hp5), 200 participants were recruited from a distance-learning university. They were offered course credit for completing the experiment. Participants were instructed to remember an incident in which they had power over another person (high-power condition) or when someone else had power over them (low-power condition). The dependent variable was the GPSPS (α = .89, ω = .89). Twenty-five individuals did not complete the power scale and/or the memory task. The final sample comprised 175 participants (22% men, 78% women; M_{age} = 32.88, SD_{age} = 10.15, 19 to 60) with 89 people in the high-power and 86 in the low-power group. Participants' memories in the recall task were rated on three categories (strong memory, weak memory, missing the point): Two independent raters assessed a subset (10%) of the memories. After establishing good interrater agreement using a quadratic weighted kappa (κ_w = .71), the remaining memories were assessed by one rater.

Results and Discussion

An independent-sample *t*-test with all participants showed a significant difference between the high-power (M = 5.04, SD = 0.99) and low-power groups (M = 4.67, SD = 1.19), $t(173) = 2.23$, p = .014, d = 0.34. When we removed participants whose narratives had been rated as "missing the point," the effect became larger (high power: M = 5.09, SD = 0.95; low power: M = 4.63, SD = 1.18), $t(155) = 2.67$, p = .004, d = 0.43. Thus, the GPSPS can be used as a state measure to assess fluctuations in people's sense of power. Such an assessment may be relevant in experimental settings or in evaluations of training, coaching, or therapy. Further, interactions of trait power with state power may be investigated in future research.

Study 5

In a final study, we wanted to further establish the validity of the state version of the GPSPS by using a different sample, a different setting (laboratory instead of online), and different power manipulation. We used the same instructions as in Study 4.

Method

The sample comprised 120 participants who were recruited at a university in southern Germany (81% women, 19% men; M_{age} = 22.56, SD_{age} = 5.86, range: 17–62 years). The students were offered course credit for completing the experiment. The power manipulation was developed in our laboratory and adapted for university students: Participants in the high-power condition were asked to imagine they lived in a large apartment and were receiving applications from potential flatmates. They had the option of choosing from among eight different applicants and were asked to figure out what they would say to applicants when interviewing them. In the low-power group, participants imagined that they had applied for a room in an apartment. They were told that they had only received a single invitation and had had a brief interview conducted in a cold manner for an unattractive room. The dependent variable was the GPSPS (α = .86, ω = .87). There were three control items about identifying with one's role in the scenario, one's motivation to work on the task, and empathizing with one's role in the scenario. Answers were given on a 7-point scale. In accordance with the preregistration, participants with a mean below 4 on the control items were excluded (https://aspredicted.org/blind.php?x=88gj7j).

Results and Discussion

First, missing values were replaced. Little's MCAR test was not significant, $\chi^2(7) = 1.529$, p = .981, which suggested that the data were missing completely at random. One missing value was replaced with the expectation-maximization method.

Then, one-tailed independent-sample *t*-tests were calculated. Results showed a significant difference between the high-power (M = 5.35, SD = 0.78) and low-power groups (M = 5.00, SD = 1.02), $t(118) = 2.14$, p = .017, d = 0.39. When we excluded participants who had a mean below 4 on the control items, the effect increased (high power: M = 5.37, SD = 0.70; low power: M = 4.97, SD = 1.04), $t(102) = 2.33$, p = .011, d = 0.45. The results suggest that the state version of the GPSPS was sensitive to an experimental power manipulation.

General Discussion

In the present studies, we analyzed the psychometric properties of the trait and state versions of the GPSPS (Anderson et al., 2012) by using five independent samples and three different instructions for the scale. With respect to the factor structure, CFAs supported a unidimensional model with six items across three studies. The two excluded items may have had different connotations for Germans compared with English-speaking participants. Corrected item-total correlations and factor loadings were high. Reliability coefficients were satisfactory in all samples, and high stability was found for the trait version of the GPSPS across three measurement occasions. The GPSPS showed strict measurement invariance across gender. With respect to nomological validity, the GPSPS was correlated with a variety of other psychological constructs in the expected direction and was thus comparable to the original scale. A personal sense of power had the strongest associations with dominance, neuroticism (negative), self-esteem, and authentic pride in the present research.

Criterion validity was established: Personal power was positively but not strongly associated with socioeconomic status. Supporting construct validity, as expected, a clinical sample scored lower on a personal sense of power than the broad sample from Study 1. Furthermore, we tested a state version to assess fluctuations in a personal sense of power. In two final studies, the state version of the GPSPS was sensitive to experimental power manipulations, but the effect sizes were rather small. Additional research will be needed to further establish the GPSPS as an adequate measure of state power. Future studies should also assess individuals' trait power and use that measure as a covariate in a subsequent experiment to better distinguish between trait and state variance.

There were no gender differences in the generalized sense of power (see Study 1), which is surprising as power is still not distributed equally between men and women in Germany (Lang & Gross, 2020). However, the assimilation of gender roles as well as increased agentic traits in women have recently been observed (Athenstaedt & Alfermann, 2011; Schwartz & Gonalons-Pons, 2016). Moreover, the generalized sense of power is an overall assessment. There is still a need to check for whether domains in which people feel powerful differ between the sexes. For example, men may report higher personal power in job-related contexts, but women might still feel more powerful in family matters (Beach & Tesser, 1993). Assessing the sense of power in different domains and testing the moderating role of sex could be a topic of future studies.

What are the theoretical implications? As the correlations in the nomological network were in the hypothesized directions for positive and negative emotions, behavioral activation, behavioral inhibition, construal style, and social distance, this provided correlational evidence in support of the approach inhibition theory of power (Keltner et al., 2003) as well as the social distance theory of power (Magee & Smith, 2013). Yet, the correlation coefficients were stronger for predictions that were based on the former theory. No associations were found between a personal sense of power and agreeableness or rivalry. The latter finding corresponds to the small positive correlation with hubristic pride and supports the notion that the experience of personal power might not be associated with antisocial attitudes but rather with high self-regard – reasoning that is in line with the high positive correlations with self-esteem, authentic pride, and narcissism. Overall, these associations are in line with theoretical assumptions and empirical findings from past power literature (Anderson & Cowan, 2014, Anderson et al., 2012).

Is a personal sense of power a cause or a consequence? Concerning the association between the GPSPS and socioeconomic status, both directions seem possible. Sociostructural power characteristics may have an impact on a personal sense of power, but a personal sense of power may also lead to high socioeconomic status. Future research should address this question in experimental and longitudinal studies. Other avenues for future research may include testing associations between a personal sense of power and gender-role self-concepts or agency versus communion and addressing the question of how experienced power varies in certain situations.

The findings in the clinical sample support the notion that personal sense of power varies with individuals' personal backgrounds. Patients with mental disorders may also benefit from interventions to increase their personal sense of power because a higher self-perceived ability to influence others and decision-making ability in interpersonal relationships are associated with desirable traits (e.g., consider the strong association between personal sense of power and emotional stability).

The project provided evidence for the unidimensionality of the scale in three independent samples. Moreover, the statistical analyses (corrected item-total correlations, reliability with different internal consistency coefficients, multigroup CFA) go beyond the analyses by Anderson et al. (2012). We used clinical, student, and community samples. Moreover, we provided evidence for the suitability of the state version of the scale. Researchers could use this scale as a manipulation check in experimental studies on power. This would be particularly promising for increasing objectivity over various power studies as researchers can directly compare their effect sizes with those of others. Such an approach would also increase the significance of statistical models with a personal sense of power as a mediator or outcome as the scale has demonstrated high reliability, and analyses would have a stronger basis.

Limitations pertain to the data sources because we used only self-reported data across the studies. Indeed, personal sense of power is a subjective assessment, but nevertheless, it would be interesting to assess self-other agreement for experienced and perceived power by using peer-report data. Another limitation is the unequal gender distribution in Studies 1, 3, 4, and 5. Women were overrepresented, which may have influenced the results of certain analyses (e.g., measurement invariance). Future research should thus aim to test the scale in samples in which men and women are represented equally. Further, it would be promising to test the scale in other interpersonal relationships (e.g., supervisor-employee) with adapted instructions. Finally, cross-cultural comparisons would be exceedingly valuable for testing whether a personal sense of power is lower or higher in certain cultures than in others and whether measurement invariance holds across cultures. Dovetailing with this issue, it is possible that a high personal sense of power in individuals from collectivistic cultures violates norms of modesty and humility and that a different pattern of correlations will thereby emerge (Morling et al., 2002). For example, there might not be a negative association between personal sense of power and negative emotions, and instead, there may be no clear correlation as the relationship may be ambiguous.

All in all, the results of the present studies provide converging evidence for the good psychometric properties of the GPSPS. We encourage researchers to use this scale as a reliable and valid instrument for assessing trait power and state power.

References

Anderson, C., & Cowan, J. (2014). Personality and status attainment: A micropolitics perspective. In J. T. Cheng, J. L. Tracy, & C. Anderson (Eds.), *The psychology of social status* (pp. 99–117). Springer.

Anderson, C., & Galinsky, A. D. (2006). Power, optimism, and risk-taking. *European Journal of Social Psychology, 36*(4), 511–536. https://doi.org/10.1002/ejsp.324

Anderson, C., John, O. P., & Keltner, D. (2012). The personal sense of power. *Journal of Personality, 80*(2), 313–344. https://doi.org/10.1111/j.1467-6494.2011.00734.x

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Athenstaedt, U., & Alfermann, D. (2011). *Geschlechterrollen und ihre Folgen* [Gender roles and their consequences]. Kohlhammer.

Back, M. D., Küfner, A. C., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. (2013). Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology, 105*(6), 1013–1037. https://doi.org/10.1037/a0034431

Beach, S. R., & Tesser, A. (1993). Decision making power and marital satisfaction: A self-evaluation maintenance perspective. *Journal of Social and Clinical Psychology, 12*(4), 471–494. https://doi.org/10.1521/jscp.1993.12.4.471

Boeker, W. (1992). Power and managerial dismissal: Scapegoating at the top. *Administrative Science Quarterly, 37*(3), 400–421. https://doi.org/10.2307/2393450

Brunell, A. B., Gentry, W. A., Campbell, W. K., Hoffman, B. J., Kuhnert, K. W., & DeMarree, K. G. (2008). Leader emergence: The case of the narcissistic leader. *Personality and Social Psychology Bulletin, 34*(12), 1663–1676. https://doi.org/10.1177/0146167208324101

Cheng, J. T., Tracy, J. L., & Henrich, J. (2010). Pride, personality, and the evolutionary foundations of human social status. *Evolution and Human Behavior, 31*(5), 334–347. https://doi.org/10.1016/j.evolhumbehav.2010.02.004

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Clauss, J. A., & Blackford, J. U. (2012). Behavioral inhibition and risk for developing social anxiety disorder: A meta-analytic study. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(10), 1066–1075. https://doi.org/10.1016/j.jaac.2012.08.002

Deuter, C. E., Schächinger, H., Best, D., & Neumann, R. (2016). Effects of two dominance manipulations on the stress response: Cognitive and embodied influences. *Biological Psychology, 119*, 184–189. https://doi.org/10.1016/j.biopsycho.2016.06.004

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment, 23*(3), 225–241. https://doi.org/10.1177/073428290502300303

Duguid, M. M., & Goncalo, J. A. (2012). Living large: The powerful overestimate their own height. *Psychological Science, 23*(1), 36–40. https://doi.org/10.1177/0956797611422915

Dunbar, N. E., & Burgoon, J. K. (2005). Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships, 22*(2), 207–233. https://doi.org/10.1177/0265407505050944

Eddington, K. M., Dolcos, F., McLean, A. N., Krishnan, K. R., Cabeza, R., & Strauman, T. J. (2009). Neural correlates of idiographic goal priming in depression: Goal-specific dysfunctions in the orbitofrontal cortex. *Social Cognitive and Affective Neuroscience, 4*(3), 238–246. https://doi.org/10.1093/scan/nsp016

Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *Journal of Personality and Social Psychology, 85*(3), 453–466. https://doi.org/10.1037/0022-3514.85.3.453

Goschke, T. (2014). Dysfunctions of decision-making and cognitive control as transdiagnostic mechanisms of mental disorders: Advances, gaps, and needs in current research. *International Journal of Methods in Psychiatric Research, 23*(S1), 41–57. https://doi.org/10.1002/mpr.1410

Guinote, A. (2017). How power affects people: Activating, wanting, and goal seeking. *Annual Review of Psychology, 68*, 353–381. https://doi.org/10.1146/annurev-psych-010416-044153

Hammervold, R. (1998). *Properties of goodness of fit statistics for structural equation models.* (Doctoral dissertation). NTNU Trondheim Norgest teknisknaturvitenskapelige Universitet.

Heineck, G. (2005). Up in the skies? The relationship between body height and earnings in Germany. *Labour, 19*(3), 469–489. https://doi.org/10.1111/j.1467-9914.2005.00302.x

Heydasch, T. (2014). *Studienerfolgsprädiktoren bei Fernstudierenden. Eine empirische Untersuchung mit Studierenden des Studiengangs B.Sc. Psychologie an der FernUniversität in Hagen* [Academic success of distance learning students. An empirical

study with BSc Psychology students at the University of Hagen]. (Dissertation thesis). University of Hagen.

Hofstee, W. K., Kiers, H. A., De Raad, B., Goldberg, L. R., & Ostendorf, F. (1997). A Comparison of Big-Five structures of personality traits in Dutch, English, and German. *European Journal of Personality, 11*(1), 15–31. https://doi.org/10.1002/(SICI)1099-0984(199703)11:1<15::AID-PER273>3.0.CO;2-8

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). https://www.InTestCom.org

Jackson, D. N. (1976). *The appraisal of personal reliability*. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Kelley, K. (2018). *The MBESS R package (version 4.4.3)*. http://cran.r-project.org

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review, 110*(2), 265–284. https://doi.org/10.1037/0033-295X.110.2.265

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kocur, D., & Mandal, E. (2018). The need for power, need for influence, sense of power, and directiveness in female and male superiors and subordinates. *Current Issues in Personality Psychology, 6*(1), 47–56. https://doi.org/10.5114/cipp.2018.72200

Körner, A., Geyer, M., Roth, M., Drapeau, M., Schmutzer, G., Albani, C., Schumann, S., & Brähler, E. (2008). Persönlichkeitsdiagnostik mit dem Neo-Fünf-Faktoren-Inventar: Die 30-Item-Kurzversion (NEO-FFI-30) [Personality assessment with the NEO Personality Inventory: 30-item short version]. *Psychotherapie Psychosomatik Medizinische Psychologie, 58*(6), 238–245. https://doi.org/10.1055/s–2007–986199

Körner, R., Heydasch, T., & Schütz, A. (2021). *GPSPS*. https://doi.org/10.17605/OSF.IO/JF9DZ

Körner, R., Petersen, L.-E., & Schütz, A. (2019). Do expansive or contractive body postures affect feelings of self-worth? High power poses impact state self-esteem. *Current Psychology*. Advance online publication. https://doi.org/10.1007/s12144-019-00371-1

Krohne, H. W., Egloff, B., Kohlmann, C., & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS) [Studies on a German version of the Positive and Negative Affect Schedule (PANAS)]. *Diagnostica, 42*(2), 139–156.

Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist, 64*(4), 241–256. https://doi.org/10.1037/a0015309

Lampert, T., Kroll, L., Müters, S., & Stolzenberg, H. (2013). Messung des sozioökonomischen Status in der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1) [Measurement of socioeconomic status in a study on adult health in Germany]. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz, 56*(5–6), 631–636. https://doi.org/10.1007/s00103-012-1663-4

Lang, V., & Gross, M. (2020). The just gender pay gap in Germany revisited: The male breadwinner model and regional differences in gender-specific role ascriptions. *Research in Social Stratification and Mobility, 65*, Article 100473. https://doi.org/10.1016/j.rssm.2020.100473

Leiner, D. J. (2013). Too fast, too straight, too weird: Post hoc identification of meaningless data in internet surveys. *SSRN Electronic Journal, 13*(3), 229–243. https://doi.org/10.18148/srm/2019.v13i3.7403

Lorr, M., & Strack, S. (1990). Wiggins interpersonal adjective scales: A dimensional view. *Personality and Individual Differences, 11*(4), 423–425. https://doi.org/10.1016/0191-8869(90)90227-I

Lysaker, P. H., Tsai, J., Yanos, P., & Roe, D. (2008). Associations of multiple domains of self-esteem with four dimensions of stigma in schizophrenia. *Schizophrenia Research, 98*(1–3), 194–200. https://doi.org/10.1016/j.schres.2007.09.035

Magee, J. C., & Smith, P. K. (2013). The social distance theory of power. *Personality and Social Psychology Review, 17*(2), 158–186. https://doi.org/10.1177/1088868312472732

Mashiach-Eizenberg, M., Hasson-Ohayon, I., Yanos, P. T., Lysaker, P. H., & Roe, D. (2013). Internalized stigma and quality of life among persons with severe mental illness: The mediating roles of self-esteem and hope. *Psychiatry Research, 208*(1), 15–20. https://doi.org/10.1016/j.psychres.2013.03.013

Morling, B., Kitayama, S., & Miyamoto, Y. (2002). Cultural practices emphasize influence in the United States and adjustment in Japan. *Personality and Social Psychology Bulletin, 28*(3), 311–323. https://doi.org/10.1177/0146167202286003

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Rost-Schaude, E., Kumpf, M., & Frey, D. (2014). Interne-Externe Kontrolle [Internal-external control]. In D. Danner & A. Glöckner-Rist (Eds.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS) Mannheim*. https://doi.org/10.6102/zis128

Russell, B. (1938). *Power: A new social analysis*. Allen and Unwin.

Schütz, A., Marcus, B., & Sellin, I. (2004). Die Messung von Narzissmus als Persönlichkeitskonstrukt [Measuring narcissism as a personality construct]. *Diagnostica, 50*(4), 202–218. https://doi.org/10.1026/0012-1924.50.4.202

Schwartz, C. R., & Gonalons-Pons, P. (2016). Trends in relative earnings and marital dissolution: Are wives who outearn their husbands still more likely to divorce? *The Russell Sage Foundation Journal of the Social Sciences, 2*(4), 218–236. https://doi.org/10.7758/RSF.2016.2.4.08

Strobel, A., Beauducel, A., Debener, S., & Brocke, B. (2001). Eine deutschsprachige Version des BIS/BAS-Fragebogens von Carver und White [A German-language version of the BIS/BAS scale by Carver and White]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 22*(3), 216–227. https://doi.org/10.1024/0170-1789.22.3.216

Stulp, G., Buunk, A. P., Verhulst, S., & Pollet, T. V. (2012). High and mighty: Height increases authority in professional refereeing. *Evolutionary Psychology, 10*(3), 588–601. https://doi.org/10.1177/147470491201000314

Tracy, J. L., & Robins, R. W. (2007). The psychological structure of pride: A tale of two facets. *Journal of Personality and Social Psychology, 92*(3), 506–525. https://doi.org/10.1037/0022-3514.92.3.506

Uziel, L., & Hefetz, U. R. I. (2014). The selfish side of self-control. *European Journal of Personality, 28*(5), 449–458. https://doi.org/10.1002/per.1972

Vallacher, R. R., & Wegner, D. M. (1989). Levels of personal agency: Individual variation in action identification. *Journal of Personality and Social Psychology, 57*(4), 660–671. https://doi.org/10.1037/0022-3514.57.4.660

Van Kleef, G. A., Oveis, C., Homan, A. C., van der Löwe, I., & Keltner, D. (2015). Power gets you high: The powerful are more inspired by themselves than by others. *Social Psychological and Personality Science, 6*(4), 472–480. https://doi.org/10.1177/1948550614566857

Von Collani, G., & Herzberg, P. Y. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg [A revised version of the German Adaptation of Rosenberg's Self-Esteem Scale]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 24*(1), 3–7. https://doi.org/10.1024/0170-1789.24.1.3

Wang, Y. N. (2015). Authenticity and relationship satisfaction: Two distinct ways of directing power to self-esteem. *PLoS One, 10*(12), e0146050. https://doi.org/10.1371/journal.pone.0146050

Weineck, F., Messner, M., Hauke, G., & Pollatos, O. (2019). Improving interoceptive ability through the practice of power posing: A pilot study. *PLoS One, 14*(2), e0211453. https://doi.org/10.1371/journal.pone.0211453

Wiggins, J. S., Trapnell, P., & Phillips, N. (1988). Psychometric and geometric characteristics of the Revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research, 23*(4), 517–530. https://doi.org/10.1207/s15327906mbr2304_8

Wojciszke, B., & Struzynska-Kujalowicz, A. (2007). Power influences self-esteem. *Social Cognition, 25*(4), 472–494. https://doi.org/10.1521/soco.2007.25.4.472

History
Received June 21, 2020
Revision received December 18, 2020
Accepted January 12, 2021
Published online April 22, 2021
EJPA Section / Category Personality

Acknowledgment
The authors are grateful to Jane Zagorski for language editing.

Conflict of Interest
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publication Ethics
All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.
Informed consent was obtained from all individual participants who were included in the study.

Open Data
Studies 1, 4, and 5 were preregistered with an analysis plan. Studies 2 and 3 were not preregistered. We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results (Körner, Heydasch, & Schütz, 2021). We also confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology (Körner, Heydasch, & Schütz, 2021). All preregistrations, data, and syntaxes for Mplus, R, and Weighted Kappa are available at https://osf.io/jf9dz/, doi: 10.17605/OSF.IO/JF9DZ.

Funding
This research was partly funded by a graduate scholarship granted by the state of Saxony-Anhalt, Germany, to Robert Körner. The source of funding had no involvement in the study design, data analysis, interpretation of data, writing, or decision to submit. Open access publication enabled by the University of Bamberg, Germany.

ORCID
Astrid Schütz
https://orcid.org/0000-0002-6358-167X

Astrid Schütz
Department of Psychology
Otto-Friedrich-University of Bamberg
Markusplatz 3
96047 Bamberg
Germany
astrid.schuetz@uni-bamberg.de

Multistudy Report

Cross-Cultural Comparison of the Benign and Malicious Envy Scale (BeMaS) Across Serbian and US Samples and Further Validation

Bojana M. Dinić[1] and Marija Branković[2]

[1]Department of Psychology, Faculty of Philosophy, University of Novi Sad, Serbia
[2]Department of Psychology, Faculty of Media and Communications, Singidunum University, Belgrade, Serbia

Abstract: The aim of this research was to validate the dual conception of envy in Serbian culture, measured by the Benign and Malicious Envy Scale (BeMaS). In Study 1 (N = 404), the results confirmed cross-cultural invariance of the Malicious Envy scale across Serbian and US samples, with the US sample obtaining higher scores. However, two items in the Benign Envy scale showed significant differential item functioning across samples. Nonetheless, both scales in Serbian showed adequate measurement precision (information) and the expected distinction in relations with narcissistic admiration, narcissistic rivalry, and self-esteem, with more aversive characteristics associated with Malicious Envy. In Study 2 (N = 404), Malicious Envy showed a negative relation with Conscientiousness and Openness, as well as higher negative correlations with Honesty-Humility, Agreeableness, psychopathy, and sadism compared to Benign Envy. Furthermore, Malicious Envy showed higher positive correlations with psychological distress, while Benign Envy showed negative correlations with some aspects of distress. The results support good psychometric properties of BeMaS scores of the Serbian adaptation and add to the cross-cultural validity of the dual conception of envy.

Keywords: Benign and Malicious Envy Scale, differential item functioning, item response theory, cross-cultural comparison, validity

Envy emerges as a result of unfavorable social comparison, in which another person is deemed superior to oneself in terms of a valued possession, quality, or achievement (Parrott & Smith, 1993). Even though envy can be viewed as an episodic emotion, there are also reasons to recognize a dispositional form of envy (Smith et al., 1999; Lange, Blatz, et al., 2018). There are several partly distinct conceptualizations of envy (see Lange, Weidman, et al., 2018), but most theorists agree on two crucial characteristics of envy. First, envy arises from upward social comparison, rendering the image of the self as inferior. Second is the psychological pain experienced due to upward social comparison, such as the feeling of inferiority (e.g., Miceli & Castelfranchi, 2007) or feelings of hostility, resentment, and hopelessness (e.g., Smith & Kim, 2007). Envy plays an important role in mental health as well as interpersonal relations. It has been related to poor mental health outcomes, including depression (Appel et al., 2015), and lowered self-esteem (Smith et al., 1999). Furthermore, envious individuals can go as far as to inflict harm on others (Duffy et al., 2012) or hurt the self or a valued object, so that the other person would not have it (Zizzo & Oswald, 2001).

However, not all authors embrace the ill-will component as inherent to envy. Instead, some emphasize the importance of the desire to level out the differences in status, which could be achieved either by leveling down the envied person or leveling up oneself (Lange & Crusius, 2015). Therefore, we will focus on this dual-facet conceptualization of envy by which it is possible to distinguish between benign and malicious envy (van de Ven et al., 2009; Lange et al., 2016). The benign form is characterized by the desire to improve oneself and emulate the envied person. The malicious form refers to what is traditionally recognized as envy and it is characterized by direct or indirect aggression toward the envied person. Both forms stem from the upward social comparison that is unfavorable to one's self-image and they both include the painful emotional component of tormenting feelings of inferiority (Lange & Crusius, 2015; Lange, Blatz, et al., 2018). This distinguishes

benign envy from positive emotions such as admiration (Lange & Crusius, 2015; van de Ven et al., 2015).

Indeed, several languages have distinct words for two forms of envy, for example, German (*beneiden* and *missgönnen,* Lange & Crusius, 2015), Polish (*zazdrość* and *zawiść*, Kwiatkowska et al., 2020) or Urdu (*rashk* and *hassad*, Khan et al., 2017). The Serbian language does not make this linguistic distinction and the used term (*zavist*) only refers to the malicious form. However, previous research has suggested that even in languages like Serbian, it is possible to discern two different emotional states related to envy (Lange, Weidman, et al., 2018).

The distinction between the two forms of envy is reflected in motivational, emotional, cognitive, and personality functioning. If the unfavorable social comparison ignites the achievement motive coupled with hope for success and a sense of personal control, the resulting emotion is benign envy. By contrast, if this motive is coupled with the fear of failure and the perception of the other's advantage as undeserved, the resulting emotion is malicious envy (Lange & Crusius, 2015; Lange et al., 2016). While experiencing both benign and malicious envy is similarly painful, previous research recognized some positive emotional components of the experience of benign envy (van de Ven et al., 2009). For instance, benign envy has been shown to be related to hope for success, a higher perception of personal control, and social potency (Lange et al., 2016). Additionally, benign envy positively predicted psychological well-being, while malicious envy negatively predicted well-being through a decreased sense of personal control (Briki, 2019). Furthermore, malicious and not benign envy predicts schadenfreude or joy at the misfortune of others (Lange, Weidman, et al., 2018; van de Ven et al., 2015). Although both forms of envy have been found to be related to narcissism, benign envy has been related to narcissistic admiration, while malicious envy has been related to narcissistic rivalry and consequently, to the propensity for social conflict (Lange et al., 2016). Malicious envy has also been uniquely related to psychopathy from the Dark Triad constellation while both benign and malicious envy have been related to Machiavellianism (Lange, Paulhus, et al., 2018). Furthermore, in a study conducted on a sample of marathon runners, benign envy was coupled with an enhanced achievement, which was mediated by higher goal setting, while malicious envy predicted goal disengagement (Lange & Crusius, 2015). These studies have established differential dynamics and performance-related outcomes of the two forms of envy.

The Present Study

Given that the two forms of envy can have different outcomes in terms of mental health outcomes and interpersonal relations, we sought to empirically validate the dual conception of envy in the Serbian culture and to establish the psychometric characteristics of the instrument constructed on the basis of this conception, namely, the Benign and Malicious Envy Scale (BeMaS; Lange & Crusius, 2015).

Given the presented theoretical and empirical rationale, we expected to validate the two forms of envy in the local cultural context. The scale has primarily been tested in the Western cultural context, for example, in Germany and the US (Lange & Crusius, 2015). Although cross-cultural validity has been established in other cultures, that is, in Japanese (Sawada & Fujii, 2016), and Turkish (Çırpan & Özdoğru, 2017), only in one study cross-cultural measurement invariance was tested across samples from Poland, Germany, Russia, and the US (Kwiatkowska et al., 2020). Since the most widely used instruments in personality and social psychology failed to provide measurement invariance across different groups (Hussey & Hughes, 2020), determining cross-cultural validity in a more precise and rigorous way seems warranted. This study was designed to contribute to the existing literature and confirm the validity of the dual conception of envy in a more collectivistic society compared to most of the previously studied countries (Hofstede, 2001), that is, in Serbian society, as well as in the context of a language that does not linguistically differentiate between the two forms of envy.

Study 1

The aim of Study 1 was to explore differential item functioning (DIF) of the BeMaS across samples from Serbia and the US as well as to test other psychometric properties of the Serbian adaptation of the BeMaS by using the Item Response Theory (IRT) analysis. We expected that the BeMaS would achieve cross-cultural invariance across Serbian and US samples and that the Serbian adaptation of the BeMaS would show good α and ω reliability coefficients and measurement precision (information) across the entire range of scale scores. Furthermore, convergent and discriminant validity of the Serbian adaptation was tested via correlations with measures of narcissism and self-esteem. Although both envy forms should correlate with narcissism, in line with previous studies, we expected that the Benign Envy scale would show a higher correlation with narcissistic admiration and a lower correlation with narcissistic rivalry, compared to the Malicious Envy scale (e.g., Lange et al., 2016). This would contribute to their discriminant validity. Moreover, since Benign Envy has been related to positive outcomes (e.g., better well-being, see Briki, 2019; hope for success and a higher sense of personal control, see Lange et al., 2016), we expected that it would show a positive correlation with self-esteem, unlike Malicious Envy.

Participants and Procedure

The Serbian sample comprised 404 students (M = 21.73, SD = 4.86), 74.8% of whom were female. The data for this study were collected within a larger cross–cultural study that aimed to determine cross–cultural validity of several instruments ("Cross-cultural study on narcissism, envy, shyness, and humor" led by researchers at the Cardinal Stefan Wyszyński University in Warsaw, Poland). There were no excluded participants or missing data. Participants were recruited among university students, in exchange for course credit. Prior to data collection, Research Ethics Board approval was obtained from the Commission of Ethics and Bioethics at Cardinal Stefan Wyszyński University in Warsaw, Poland (registration number: KEiB – 14/2017).

The US sample was extracted from Lange, Paulhus, et al. (2018), whereby 5 MTurk samples (https://osf.io/mb74v/) in this study were merged and selected subsample which matches the Serbian sample in both sample size and age (the upper third of the total sample). Only those who reported that English was their mother tongue were included. The extracted sample comprised 417 participants (41.7% females).

Instruments

The Benign and Malicious Envy Scale (BeMaS; Lange & Crusius, 2015, for the Serbian adaptation see https://osf.io/3msne/) consists of 10 items with a 6-point Likert scale (from 1 = *strongly disagree* to 6 = *strongly agree*). Five items refer to Benign Envy and the remaining 5 refer to Malicious Envy. Besides the BeMaS, two more measures were used: (1) a short-version of the Narcissistic Admiration and Rivalry Questionnaire (NARQ-S; Back et al., 2013, for the Serbian adaptation of the long-version of the scale see Gojković et al., 2019), which comprises 6 items with a 6-point Likert scale (from 1 = *not agree at all* to 6 = *agree completely*), of which 3 items measure Narcissistic Admiration (α = .80) and the remaining 3 measure Narcissistic Rivalry (α = .56). Since this is the first use of NARQ-S, model fit in this study was good: CFI = .98, TLI = .96, RMSEA = .07, 90% CI [.03, .10], SRMR = .04, and better than one-factor solution ($\Delta\chi^2(1) = 48.17$, $p < .001$); (2) the Single-Item Self-Esteem Scale (Robins et al., 2001), which contains one item with a 7-point Likert scale (from 1 = *not very true of me* to 7 = *very true of me*).

Data Analysis

First, a confirmatory factor analysis (CFA) with a maximum likelihood estimator was conducted in order to test a two-factor model of the BeMaS ("lavaan" R package; Rosseel, 2012). In line with recommendations required indices for an excellent fit are RMSEA and SRMR < .06, and TLI and CFI > .95, and for an acceptable fit are RMSEA and SRMR < .08, and TLI and CFI > .89 (Greiff & Allen, 2018). A substantive convergent validity is achieved when all item loadings are significant and the average variance extracted (AVE; see Fornell & Larcker, 1981) is higher than .50 within each factor.

Second, the IRT analysis was applied in order to: (1) detect DIF between Serbian and US samples, and (2) test the psychometric properties of the Serbian version. Items flagged for DIF indicated that participants of two samples, who have equal levels of the latent trait, do not have the same probability of endorsing the item. There are two types of DIF: (1) uniform, in which DIF effect remains constant across the continuum of the latent trait, and (2) nonuniform, in which the strength or direction of the DIF effect is not the same across the continuum of the latent trait. Change higher than 0.02 in McFadden's pseudo R^2 indicated significant DIF ("lordif" R package, Choi et al., 2011). In addition to DIF, differential test functioning (DTF) was also calculated to assess the impact of DIF on the total scale score. DTF was calculated via an analysis of covariance in which the sample (Serbian or the US) was entered as a factor, the average score only on DIF-free items as the covariate, and the total average score on all items as the dependent variable. The resulting difference in mean total scores between samples was then divided by the standard deviation of the US group to obtain effect size (d_{DTF}). This effect size was interpreted in accordance with Cohen's (1988) rule of thumb: 0.2 for a small effect, 0.5 for a moderate effect, and 0.8 or higher for a large effect.

Furthermore, the IRT graded response model was conducted in "ltm" R package (Rizopoulos, 2006). Two item parameters were analyzed: difficulty (β), which refers to the amount of the latent trait necessary to have a 50% chance of endorsing the item, and discrimination (a), which indicated how well an item can differentiate between participants at different trait levels. Discrimination parameters up to 0.64 are low, those between 0.65 and 1.34 are moderate, those between 1.35 and 1.69 are high, and those over 1.7 are very high (Baker, 2001). The key characteristic in the IRT is information, which reflects measurement reliability or precision at each level of the latent trait. Prior the main IRT analysis, the unidimensionality of each scale was tested via parallel analysis, and absence of misfit combinations of items.

Third, convergent and discriminant validity correlations with narcissism and self-esteem measures were calculated with Steiger's Z test for testing the significance of dependent correlations.

The sample size was determined in line with recommendations that for multi-group modeling the rule of thumb is a minimum of 100 cases/observations per group (Kline, 2016).

Data and R code for both studies are available at https://osf.io/3msne/.

Design and Analysis Transparency Statement

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact p values, effect sizes, and 95% confidence or credible intervals.

Results

The results of the CFA showed that the US version had excellent CFI, TLI, and SRMR indices, while RMSEA was acceptable ($\chi^2(34) = 116.03$, CFI = .96, TLI = .95, RMSEA = .08, 90% CI [.06, .09], SRMR = .05). The Serbian version had excellent CFI and TLI, acceptable SRMR, and questionable RMSEA ($\chi^2(34) = 137.08$, CFI = .95, TLI = .93, RMSEA = .09, 90% CI [.07, .10], SRMR = .08). Overall, the majority of the indices for the Serbian version had acceptable model fit. One-factor model was included for model comparison in line with other conceptualizations of envy (e.g., Miceli & Castelfranchi, 2007). Results showed that the two-factor model was better than the one-factor model in both samples (Serbian: $\Delta\chi^2(1) = 540.69$, $p < .001$; the US: $\Delta\chi^2(1) = 1237.4$, $p < .001$). All loadings on the Serbian version were significant and high (ranged from .55 to .84) and the correlation between the two factors was significant although low ($r = .34$, $p < .001$). Both scales of the Serbian adaptation showed high AVE values (Benign Envy: AVE = .51; Malicious Envy: AVE = .57), which confirmed their convergent validity (Table 1). Reliability based on Cronbach's α and McDonald's ω coefficients was good for both scales (Table 1). There were no sex differences in either of the scales (Benign Envy: $t(402) = -0.60$, $p = .551$, $d = 0.07$; Malicious Envy: $t(402) = -0.40$, $p = .689$, $d = 0.05$).

Both scales showed unidimensionality (see Figure A in the supplementary material available at https://osf.io/3msne/) and correlations between the residuals were small within the scales, ranged from −.15 to .11. There were no flagged two-way or three-way misfit combinations of items (see Table A in the supplementary material available at https://osf.io/3msne/). The DIF analysis on the Benign Envy scale resulted in 2 items flagged for DIF (Figure 1). Both flagged items showed uniform DIF. Discrimination parameters were higher in the US sample. Item response functions suggested that category threshold parameters for the US sample were uniformly smaller than those for the Serbian sample (Figure 1). Thus, participants in the US sample were more likely to endorse these items. Effect size for the DTF effect was moderate ($d_{DTF} = 0.47$). Thus, it can be concluded that responses on these two items are culturally specific, which precludes comparison between the samples.

No items on the Malicious Envy scale were flagged for DIF. Thus, a comparison including scores on this scale was justified. The results showed that the US sample had higher malicious envy compared to the Serbian sample, with a large effect size ($t(819) = -15.30$, $p < .001$, $M_{Diff} = -1.11$, 95% CI [−1.25, −0.97], $d = 1.07$). In order to check whether sex influenced the obtained differences due to the unbalanced sex distribution across cultural groups, an additional two-way factor ANOVA was conducted with sex and culture as factors. The results showed that there was neither a significant effect of sex, $F(1, 817) = 1.61$, $p = .205$, nor a significant interaction between sex and culture, $F(1, 817) = 0.60$, $p = .440$. Thus, differences in Malicious Envy could be attributed to the effect of the culture.

The IRT analysis on the Serbian adaption of the BeMaS showed that two items (B3 and M1) had high discrimination parameters, while the rest of the items had very high discrimination parameters (Table 2). Benign Envy items adequately discriminated among people along with the whole trait range, while Malicious Envy items were more "difficult" to endorse (e.g., for choosing category "1," the average level of the trait is needed). It should be noted that items flagged for DIF (B3 and B4) were the most difficult and had the lowest (although still high according to cut-off values) discrimination parameter in the Benign Envy scale.

The IRT analysis showed good information on both scales, with the Benign Envy scale being most informative in the range of average scores and Malicious Envy in the range of above-average scores (Figure 2).

Furthermore, correlations with narcissism dimensions showed that Benign Envy correlated higher with admiration and lower with rivalry compared to Malicious Envy, even after controlling for the shared variance among Benign Envy and Malicious Envy scales (Table 3). The correlation with Narcissistic Rivalry was still higher for Malicious Envy (.79) compared to Benign Envy (.40, Steiger's $Z = 10.32$, $p < .001$) after the correction for low reliability of Narcissistic Rivalry. Moreover, BE correlated positively with self-esteem whereas Malicious Envy correlated negatively with it, although both correlations were among the lowest. The same pattern remained after controlling for the shared variance, but the correlations were somewhat higher.

Table 1. Descriptives and reliabilities for the BeMaS in Serbian and US samples

	Serbia (n = 404)				US (n = 417)			
	M	*SD*	α	Ω	*M*	*SD*	α	ω
Benign envy	3.43	1.28	.83	.84	4.17	1.06	.86	.86
Malicious envy	1.58	0.85	.87	.87	2.69	1.19	.90	.90

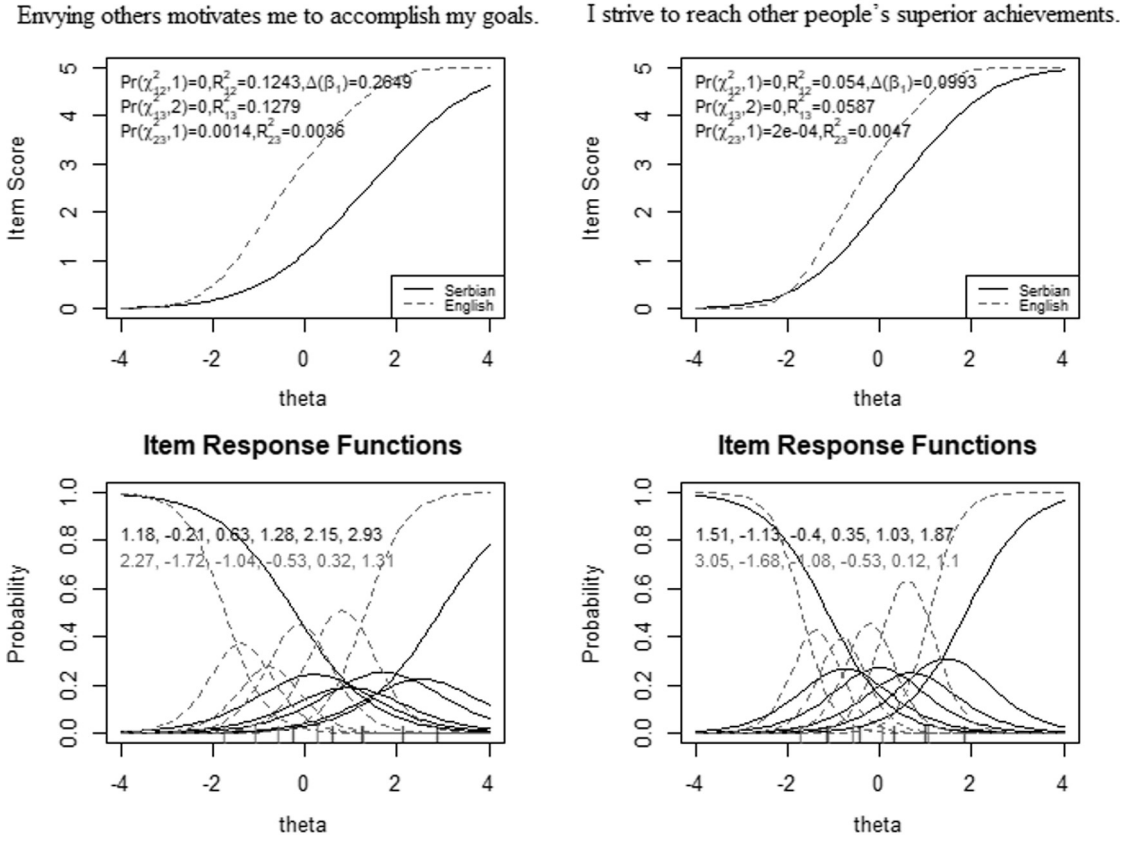


Figure 1. Benign Envy items with differential item functioning across Serbian and US samples.

Table 2. Item Response Theory parameters of items of the Serbian adaptation of the BeMaS

Item code	No. in BeMaS	β_1	β_2	β_3	β_4	β_5	α
B1	1	−1.37	−0.82	−0.41	0.14	0.98	2.07
B2	3	−1.82	−1.34	−0.72	−0.08	0.71	2.44
B3	4	−0.15	0.57	1.12	1.87	2.53	1.40
B4	7	−0.91	−0.30	0.30	0.88	1.59	1.86
B5	9	−1.06	−0.60	−0.16	0.36	0.84	3.63
M1	2	−0.17	0.83	1.56	2.58	3.37	1.43
M2	5	0.73	1.33	1.82	2.30	2.53	3.18
M3	6	0.51	1.08	1.45	1.97	2.29	3.57
M4	8	0.79	1.42	1.69	2.04	2.66	3.50
M5	10	0.70	1.39	1.81	2.41	2.94	3.05

Note. β_{1-4} = item difficulty parameter for each response category, α = discrimination parameter.

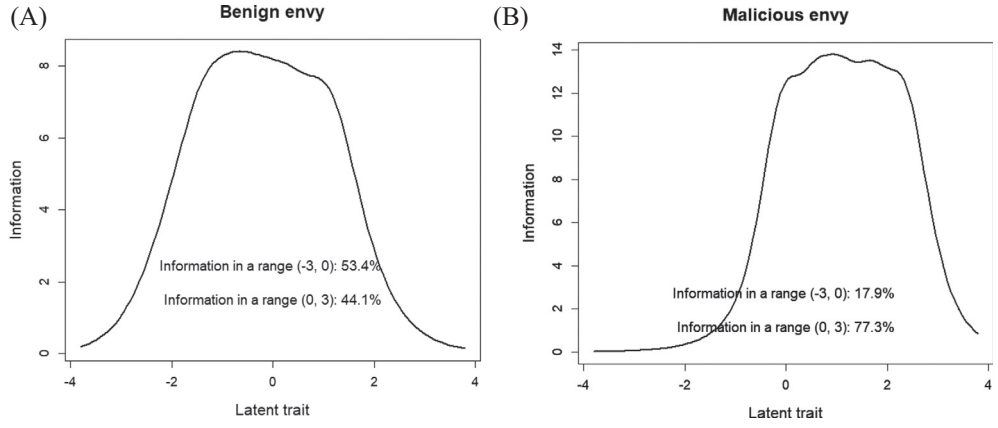


Figure 2. Information of the Serbian adaptation of Benign Envy (A) and Malicious Envy (B) scales.

Table 3. Zero-order and partial correlations between the Serbian adaptation of the BeMaS and narcissism and self-esteem measures

	Benign envy	Malicious envy	Steiger's *Z*
Malicious envy	.35***	1	–
Narcissistic admiration	.28*** (.24***)	.17** (.08)	0.21*
Narcissistic rivalry	.27*** (.10*)	.55*** (.51***)	−5.68***
Self-esteem	.12* (.21***)	−.21*** (−.27***)	5.87***

Note. Presented in the parentheses are partial correlations for BeMaS scales, after controlling for the other scale from the BeMaS. Steiger's *Z* was calculated on zero-order correlations. ***$p < .001$; **$p < .01$; *$p < .05$.

Discussion

The results of Study 1 supported the proposed two-factor solution of the Serbian adaptation of the BeMaS, with low correlations among factors. Both Benign Envy and Malicious Envy scales had a good internal consistency, which is in line with previous studies (e.g., Lange & Crusius, 2015). The results of the IRT analysis supported good measurement precision of both scales. However, it was noticeable that the Malicious Envy scale was more precise in the above-average score range. Thus, it seems more appropriate for those who could manifest this more socially aversive form of envy. This phenomenon commonly occurs with measures of socially undesirable traits (e.g., Dinić et al., 2018). Almost all items of the Malicious Envy scale were difficult to endorse, which affected measurement precision at lower trait levels.

Furthermore, the results of the DIF analysis showed that two items from the Benign Envy scale were not cross-culturally invariant ("Envying others motivates me to accomplish my goals" and "I strive to reach other people's superior achievements"). These two items seem to be more general and do not include direct, explicit comparison with another person, but rather a general feeling of envy as a source of motivation. By comparison, other items from the Benign Envy scale included direct comparison with others and a direct source of perceived threat (e.g., "When I envy others…"; "If I notice that another person is better than me…").

On the other hand, the Malicious Envy scale achieved cross-cultural invariance. Participants from the US showed higher Malicious Envy scores compared to participants from Serbia. While the mainstream American culture is characterized by high individualism and orientation toward self, Serbian culture is more collectivistic (Hofstede, 2001). Thus, differences in malicious envy could indicate that individualistic cultures prioritize personal benefits over group benefits, coupled with a more competitive environment. It should be mentioned that in only one study, measurement invariance of the BeMaS was tested across samples from the US, Germany, Poland, and Russia and results showed that the largest number of non-invariant parameters concerned the Polish sample (Kwiatkowska et al., 2020). However, additional analysis showed that the scale could be considered as invariant since less than 25% of parameters were non-invariant.

To sum up, the two forms of envy, malicious and benign, were confirmed in the Serbian culture, despite the fact that there is no linguistic distinction between these forms. However, only the Malicious Envy scale showed cross-cultural invariance. Thus, participants from the Serbian sample obtained lower Malicious Envy scores, which could reflect the different social norms in these two cultures.

In line with some previous studies, both forms of envy were found to be positively related to narcissism (Lange et al., 2016). However, Benign Envy was more strongly related to Narcissistic Admiration, consistent with perceiving the envied person as socially potent, while Malicious Envy was more strongly related to Narcissistic Rivalry, which implies a clearer propensity for social conflict (Lange et al., 2016). Moreover, Malicious Envy showed a higher and negative relation to self-esteem compared to Benign Envy, which showed the opposite direction of the relation. This adds to previous literature on the relationship between envy and self-esteem (Smith et al., 1999), specifying that this relation depends on the type of envy experienced. Thus, correlations with narcissism and self-esteem confirmed the convergent and discriminative validity of the Serbian adaptation of the BeMaS, indicating a more aversive nature of malicious envy.

Study 2

The aim of Study 2 was to further validate the Serbian adaptation of the BeMaS. In previous studies, Benign Envy was associated with Machiavellianism and to a lesser extent with grandiose narcissism, while Malicious Envy was associated with both Machiavellianism and psychopathy from the Dark Triad (e.g., Lange, Paulhus, et al., 2018). Among HEXACO traits, Honesty-Humility could be seen as the "core" element of the Dark Tetrad (the Dark Triad + sadism, e.g., Book et al., 2016). Previous studies have shown that both forms of envy predict morally questionable behaviors (see Crusius et al., 2020), including those whose predisposition is Honesty-Humility, such as deception and manipulative interpersonal behavior. Thus, for convergent validity, we expected that both envy scales would show strong negative correlations with Honesty-Humility and positive with dark traits. In line with previous studies (e.g., Lange, Paulhus, et al., 2018), correlations with Malicious Envy should be higher for these traits, with the exception of narcissism, which should contribute to the discriminant validity of the two envy forms. For further testing of the discriminant validity, we expected that correlations with the remaining HEXACO traits would be lower compared to correlations with Honesty-Humility and Dark Tetrad traits. Moreover, since Malicious Envy is characterized by low self-control (e.g., Briki, 2019; Crusius et al., 2020), we expect that it would show a stronger negative correlation with Conscientiousness, compared to Benign Envy. Additionally, as Malicious Envy involves hostile feelings toward superior others (e.g., Crusius et al., 2020), we expect that it would negatively correlate with Agreeableness, which contains hostility, anger, and impatience on its negative pole in the HEXACO model (e.g., Ashton & Lee, 2009). Criterion validity was further tested by establishing correlations with aspects of psychological distress. In a recent review study, Crusius et al. (2020) highlighted that both aspects of envy could be functional or dysfunctional, depending on the context. However, when general self-report measures were included without any experimental manipulation, previous studies showed that Benign Envy positively predicted well-being while the opposite was true for Malicious Envy (Briki, 2019). Therefore, we expected that both Benign Envy and Malicious Envy scales would significantly correlate with psychological distress, although Malicious Envy should be more strongly related to indicators of distress.

Participants and Procedure

The sample comprised 404 participants (49.5% males) from the general population from Serbia, aged between 20 and 76 years (M = 34.59, SD = 11.95), of whom 32.2% had finished high school, 29% were university students, 12.6% had finished college, and 26.2% had a university degree. The sample was collected by trained undergraduate students as a part of their pre-exam activity. In order to collect data from a heterogeneous sample, each student collected data from six participants, in accordance with the given gender and age quotas (three age groups: 20–29, 30–39, 40 years and older, with both male and female participants in each age group). The data for this study were collected within a larger study, which also contained data for other instruments. The study of Dinić, Sadiković, et al. (2020) was conducted from the same dataset, but with different instruments and aims. There were no excluded participants or missing data. The study was approved by the Ethical Committee of the Department of Psychology, Faculty of Philosophy, University of Novi Sad, Serbia, which is the Second Instance Commission of the Ethical Committee of the Serbian Psychological Society.

Instruments

Five instruments were administered:

(1) The BeMaS;
(2) The HEXACO-60 (Ashton & Lee, 2009, for the Serbian adaptation of the long version see Međedović et al., 2019, and for short see, for example, Dinić et al., 2018), which is a 60-item measure of six traits from the lexical HEXACO model of personality: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience;
(3) The Short Dark Triad (SD3; Jones & Paulhus, 2014, for the Serbian adaptation see Dinić et al., 2018), which measures three dark traits, that is, the Dark Triad (Machiavellianism, narcissism, and psychopathy), with 9 items per trait;

(4) The Short Sadistic Impulse Scale (SSIS; O'Meara et al., 2011, for the Serbian adaptation see Dinić, Bulut Allred, et al., 2020), which contains 10 items and measures sadism as the fourth dark trait;
(5) The Clinical Outcomes in Routine Evaluation – Outcome Measure CORE-OM (Evans et al., 2000), which contains 34 items measuring four aspects of psychological distress – (poor) subjective well-being (4 items), problems and symptoms including anxiety, depression, somatic symptoms, and the like (12), (poor) functioning, including general functioning and functioning in close and social relationships (12), and risk, including harm to self and harm to others (6), with higher scores corresponding to higher psychological distress. Due to the similarity between the Serbian and Croatian languages, an already established Croatian translation (Jokić-Begić et al., 2014) was adapted to the Serbian language. For the Serbian adaptation, see Dinić, Sadiković, et al. (2020) in which the same dataset was used but with other aims and sets of instruments. All measures contain a 5-point Likert-type scale for answering. Cronbach's αs and ωs are presented in Table 3.

Data Analysis

First, a CFA was conducted in order to check the model fit on this sample. To determine the model fit, the same criteria were used as in Study 1 (see Greiff & Allen, 2018). For the minimum sample size for CFA, we followed the recommendation of $N = 200$ (Kline, 2016). Second, convergent and discriminant validity correlations with used measures were calculated, with Steiger's Z test for testing the significance of dependent correlations. Profile similarity between the two scales was calculated by Cronbach and Gleser's (1953) D statistics based on Euclidean distances. Lower values indicated greater profile similarity and D could be interpreted as Cohen's d. The value of 0.41 was interpreted as the minimum effect size representing a "practically" significant effect for social science data (Ferguson, 2009).

Design and Analysis Transparency Statement

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact p values, effect sizes, and 95% confidence or credible intervals.

Results

The results of the CFA confirmed that the two-factor model was better than the one-factor model ($\Delta\chi^2(1) = 583.45$, $p < .001$). In the two-factor model solution, CFI and TLI showed good model fit, SRMR questionable, and RMSEA poor ($\chi^2(34) = 167.23$, CFI = .92, TLI = .90, RMSEA = .10, 95% CI [.08, .11], SRMR = .09). Correlations between the residuals were small within the scales, ranged from −.05 to .04. The highest modification indices include the B3 item which had significant DIF in Study 1. However, considering that only one fit index (RMSEA) showed poor model fit, we kept the original two-factor model in further analyses. All loadings were high, ranging from .69 to .85, with a significant but low correlation between factors ($r = .33$, $p < .001$). Moreover, AVE was .53 for Benign Envy and .49 for Malicious Envy, indicating adequate convergent validity.

Cronbach's αs for BeMaS scales were good (Table 4) and McDonald's ω coefficients were .84 for Benign Envy and .83 for Malicious Envy scales. Compared to Benign Envy, Malicious Envy showed higher negative correlations with Honesty-Humility, Agreeableness, Conscientiousness, and Openness and higher positive correlations with psychopathy and sadism. Regarding relations with psychological distress, Malicious Envy showed higher positive correlations with psychopathological problems/symptoms, general and interpersonal functioning, and risky behaviors. In a similar vein, Benign Envy showed a higher negative correlation with poor well-being, indicating that better well-being was related to benign envy. Partial correlations mostly showed the same relationship pattern, with some exceptions. First, the Benign Envy scale showed a significant negative correlation with sadism and a significant positive correlation with Conscientiousness when the shared variance with the Malicious Envy scale was controlled. Second, the Malicious Envy scale showed a significant positive correlation with poor well-being. All these correlations were small. Profile similarity between Benign Envy and Malicious Envy scales was .91, which indicated a large distinction between the scales.

Discussion

The results of Study 2 add further support to the validation of the Serbian adaptation of the BeMaS based on the distinction between the two forms of envy which showed large profile dissimilarity. Compared to Benign Envy, the Malicious Envy scale showed significantly higher negative correlations with Conscientiousness and Openness, which indicated that impulsivity and rigid behavioral patterns were related to malicious envy. Other studies have also shown that a lack of self-control is a specific correlate of malicious but not benign envy (e.g., Briki, 2019).

In line with previous findings (e.g., Lange, Paulhus, et al., 2018), both forms of envy were positively related to dark traits. The Malicious Envy scale showed higher correlations

Table 4. Descriptives, Cronbach's α, and validity zero-order and partial correlations of the Serbian adaptation of the BeMaS

	Benign envy	Malicious envy	Steiger's *Z*	*M*	*SD*	α
Benign envy	1			3.05	0.98	.84
Malicious envy	.33***	1	–	1.67	0.74	.83
Honesty-Humility	−.36*** (−.25***)	−.47*** (−.40***)	2.17*	3.40	0.65	.71
Emotionality	−.04 (−.01)	−.08 (−.07)	0.69	3.11	0.65	.75
Extraversion	.16*** (.21***)	−.10* (−.17***)	1.05	3.36	0.63	.79
Agreeableness	−.22*** (−.13**)	−.33*** (−.28***)	2.01*	3.09	0.57	.71
Conscientiousness	.09 (.17***)	−.21*** (−.25***)	5.24***	3.66	0.61	.78
Openness to experience	−.02 (.06)	−.22*** (−.23***)	3.51***	3.43	0.75	.80
Machiavellianism	.35*** (.25***)	.43*** (.36***)	−1.55	2.91	0.69	.80
Narcissism	.36*** (.27***)	.40*** (.32***)	−0.77	2.68	0.65	.73
Psychopathy	.32*** (.17***)	.58*** (.53***)	−5.33***	2.01	0.63	.74
Sadism	.12* (−.10*)	.59*** (.59***)	−9.18***	1.36	0.55	.86
Poor well-being	−.15** (−.19***)	.08 (.14**)	−4.00***	2.38	0.71	.69
Symptoms	−.02 (−.07)	.14** (.15**)	−2.78**	2.39	0.74	.90
Functioning	−.10* (−.20***)	.24*** (.30***)	−5.96***	2.12	0.56	.82
Risk	.04 (−.08)	.35*** (.36***)	−5.55***	1.26	0.49	.83

Note. Presented in the parentheses are partial correlations for BeMaS scales, after controlling for the other scale from the BeMaS. Steiger's *Z* was calculated on zero-order correlations. A part of the data was used in Dinić, Sadiković et al. (2020). ***$p < .001$; **$p < .01$; *$p < .05$.

with psychopathy and sadism, highlighting malevolent characteristics of malicious envy. This became more obvious in partial correlations between Benign Envy and sadism, which were negative. Benign Envy also showed malevolent characteristics, but not as prominently as Malicious Envy. Given that both envy forms positively correlate with Dark Triad traits, negative correlations with basic traits related to antagonism, Honesty-Humility, and Agreeableness were expected. The results showed that the Malicious Envy scale had somewhat higher negative correlations with these two HEXACO traits, compared to the Benign Envy scale, which further supports the malevolent nature of malicious envy.

Considering relations with psychological distress domains, it could be seen that the Malicious Envy scale was related to indicators of psychopathological symptoms, impaired functioning, and interpersonal problems as well as with aggressive behaviors toward others and self. Thus, Malicious Envy is associated with more distress and poorer functioning, in general. Partial correlations further support this conclusion. On the other hand, although the Benign Envy scale is associated with aversive traits, it was related to better well-being, which is in line with previous studies that investigated associations with well-being (e.g., Briki, 2019) as well as with studies showing positive relations between benign envy and positive emotional and motivational states (e.g., Lange & Crusius, 2015; Lange et al., 2016). However, it should be highlighted that from a functional standpoint, both aspects of envy represent reactions to threat that contain different self-defensive strategies and that both benign and malicious envy has a "dark" side and could lead to maladaptive outcomes (see Crusius et al., 2020).

Interestingly, Emotionality from HEXACO did not significantly correlate with either envy form. This result indicates the conceptualization of Emotionality in the HEXACO model, which includes anxiety and fearfulness as the common indicators of Neuroticism, but not as anger-related indicators (e.g., Ashton & Lee, 2009). In fact, those indicators are placed in the Agreeableness domain of the HEXACO model. Previous research has also found no significant correlation between envy and HEXACO Emotionality (Wilkin & Connelly, 2015).

To sum up, the findings of Study 2 showed distinct correlates of the two forms of envy, with malicious envy being related to more malevolent characteristics, impulsivity-related behaviors, and more psychological distress. Although benign envy also showed malevolent characteristics, it was related to less psychological distress. The results add to cross-cultural validity of the BeMaS scale and support its usefulness in the local context.

General Discussion

The aim of this multi-study research was to explore the psychometric characteristics of the Serbian adaptation of the BeMaS and to provide further evidence of the scale's cross-cultural validity. Given the problem with "hidden" invalidity among psychological instruments, including failed measurement invariance (Hussey & Hughes, 2020), this research offers sophisticated and rigorous tests of the

cross-cultural validity of the BeMaS. The results showed that the Benign Envy scale, or more precisely, two items from this scale, seemed biased, with higher endorsement among the US participants. In addition, one of these two items (B3) also contributed to the somewhat lower model fit which suggested that it could be revised. This is the most difficult item in the Benign Envy scale and the only one which does not include an explicit, direct comparison between envied person and the person who envies (or his/her achievements). By contrast, the Malicious Envy scale was invariant across Serbian and US samples, which allowed for a comparison between these samples. The results showed higher Malicious Envy scores in the US sample, compared to the Serbian sample. A previous study showed significantly higher scores on both envy scales in US participants compared to German, Polish, and Russian participants (Kwiatkowska et al., 2020). In more individualistic cultures, competition and outperforming others is seen as a desirable aspect of social relations. Thus, individuals who live in such cultures derive pleasure from being envied and expect others to suffer more from not having what they desire (Mosquera et al., 2010).

The results across the two large sample studies showed that the Serbian adaptation of the BeMaS is a reliable and valid measure, with a clear distinction between the two envy forms. Similar to previous validations, the confirmatory factor analyses clearly favored the two-factor solution over the single-factor model (e.g., Kwiatkowska et al., 2020). Furthermore, across both studies, Malicious Envy and Benign Envy scales showed a partly differential pattern of relations with personality traits, as well as various indicators of mental health, which is in line with previous studies (Lange et al., 2016; Lange, Weidman, et al., 2018). Malicious Envy in specific was significantly more related to morality issues, a lack of control, rigid behavioral patterns, and malevolence toward others as well as an impaired self-esteem and psychological distress. On the other hand, even if related to dark traits, Benign Envy appears to mitigate some of the problematic characteristics of the malicious form: poor well-being and problems in general and social functioning. Thus, benign envy appears to be associated with less experienced distress in the domain of mental health. Social ties are less negatively affected by benign envy, which can also present a positive motivational influence on the person who experiences it. Since benign envy entails a higher sense of control (Briki, 2019) and hopes for success (Lange et al., 2016), the psychological pain instigated by this form of envy could be more tolerable. Further research is needed to more clearly establish the crucial ingredients that differentiate the experience of malicious and benign envy.

The present study has some limitations. As our study was cross-sectional, it did not allow for any conclusions about causal relations. In self-report questionnaires that measure aversive traits, social desirability is always a potential problem. Previous research about relations between the Dark Triad traits and social desirability has suggested that individuals scoring higher on more antagonistic traits are less concerned with social desirability (Kowalski et al., 2018). Thus, the same could be expected for those who scored higher on both BeMaS scales, but future studies should address this potential issue. Next, not all model fit indices were acceptable for the Serbian adaptation of the BeMaS. Hence, future studies should consider reformulating some items to suit the Serbian cultural context, especially items that showed significant DIF. Additionally, although we provided α and ω reliability coefficients, there is no test for other types of reliability such as test-retest reliability. Moreover, we investigated only some correlates of the two envy forms to establish the basic validity of the Serbian BeMaS. Future studies could expand the nomological network, in particular within the domain of interpersonal and social relations, which appears to be closely affected by the different experiences of envy.

Taken together, the results showed the expected factor structure, good internal consistency and information, as well as convergent, discriminant, and criterion validity of the scores of the Serbian adaptation of the BeMaS. Benign Envy and Malicious Envy scales are related to relatively distinct experiences and partly different personal and interpersonal outcomes. The results supported the dual conception of envy in Serbian culture and the Serbian language, which does not have a unique term for each of the two envy forms. Given the high prevalence of social comparison situations and the resulting envy in everyday life (Foster et al., 1972), the dual model of envy could afford a better understanding of cultural varieties of experiences and outcomes of envy.

References

Appel, H., Crusius, J., & Gerlach, A. L. (2015). Social comparison, envy, and depression on Facebook: A study looking at the effects of high comparison standards on depressed individuals. *Journal of Social and Clinical Psychology, 34*(4), 277–289. https://doi.org/10.1521/jscp.2015.34.4.277

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Back, M. D., Küfner, A. C. P., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. A. (2013). Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology, 105*(6), 1013–1037. https://doi.org/10.1037/A0034431

Baker, F. B. (2001). *The basics of Item Response Theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Book, A., Visser, B. A., Blais, J., Hosker-Field, A., Methot-Jones, T., Gauthier, N. Y., Volka, A., Holden, R. R., & D'Agata, M. T. (2016).

Unpacking more "evil": What is at the core of the dark tetrad? *Personality and Individual Differences, 90*, 269–272. https://doi.org/10.1016/j.paid.2015.11.009

Briki, W. (2019). Harmed trait self–control: Why do people with a higher dispositional Malicious envy experience lower subjective wellbeing? A cross-sectional study. *Journal of Happiness Studies, 20*(2), 523–540. https://doi.org/10.1007/s10902-017-9955-x

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/Item Response Theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1–30. https://doi.org/10.18637/jss.v039.i08

Çırpan, Y., & Özdoğru, A. A. (2017). Turkish adaptation of BeMaS Benign and Malicious Envy Scale: Transliteral equivalence, reliability and validity study. *Anatolian Journal of Psychiatry, 18*(6), 577–585. https://doi.org/10.5455/apd.256664

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*(6), 456–473. https://doi.org/10.1037/h0057173

Crusius, J., Gonzalez, M. F., Lange, J., & Cohen-Charash, Y. (2020). Envy: An adversarial review and comparison of two competing views. *Emotion Review, 12*(1), 3–21. https://doi.org/10.1177/1754073919873131

Dinić, B. M., & Branković, M. (2021). *Cross-cultural comparison of the Benign and Malicious Envy Scale (BeMaS) across Serbian and US samples and further validation*. https://osf.io/3msne/

Dinić, B. M., Bulut Allred, T., Petrović, B., & Wertag, A. (2020). A test of three sadism measures: Short Sadistic Impulse Scale, varieties of sadistic tendencies, and assessment of sadistic personality. *Journal of Individual Differences, 41*(4), 219–227. https://doi.org/10.1027/1614-0001/a000319

Dinić, B. M., Petrović, B., & Jonason, P. K. (2018). Serbian adaptations of the Dark Triad Dirty Dozen (DTDD) and Short Dark Triad (SD3). *Personality and Individual Differences, 134*, 321–328. https://doi.org/10.1016/j.paid.2018.06.018

Dinić, B. M., Sadiković, S., & Wertag, A. (2020). Factor mixture analysis of the Dark Triad and Dark Tetrad. Could sadism make a difference? *Journal of Individual Difference*. Advance online publication. https://doi.org/10.1027/1614-0001/a000331

Duffy, M. K., Scott, K. L., Shaw, J. D., Tepper, B. J., & Aquino, K. (2012). A social context model of envy and social undermining. *Academy of Management Journal, 55*(3), 643–666. https://doi.org/10.5465/amj.2009.0804

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clinical outcomes in routine evaluation. *Journal of Mental Health, 9*(3), 247–255. https://doi.org/10.1080/713680250

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538. https://doi.org/10.1037/a0015808

Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.2307/3151312

Foster, G. M., Apthorpe, R. J., Bernard, H. R., Bock, B., Brogger, J., Brown, J. K., Cappannari, S. C., Cuisenier, J., D'Andrade Roy, G., Faris, J., Freeman, S. T., Kolenda, P., MacCoby, M., Messing, S. D., Moreno-Navarro, I., Paddock, J., Reynolds, H. R., Ritchie, J., Erlich, V., . . . Whiting, B. (1972). The anatomy of envy: A study in symbolic behavior [and comments and reply]. *Current Anthropology, 13*(2), 165–202.

Gojković, V., Plahota, M., & Dostanić, J. (2019). Narcizam SD3 i narcizam modela NARC: Sličnosti i razlike [The SD3 measure of narcissism and the narcissism of the NARC model: Differences and similarities]. *Zbornik Instituta za kriminološka i sociološka istraživanja, 38*(3), 25–45.

Greiff, S., & Allen, M. S. (2018). EJPA Introduces Registered Reports as New Submission Format. *European Journal of Psychological Assessment, 34*(1), 217–219. https://doi.org/10.1027/1015-5759/a000492

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage Publications.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Jokić-Begić, N., Lauri Korajlija, A., & Jurin, T. (2014). Faktorska struktura, psihometrijske karakteristike i kritična vrijednost hrvatskog prevoda CORE-OM upitnika [Factor structure, psychometric properties and cut-off scores of Croatian version of Clinical Outcomes in Routines Evaluation – Outcome Measure (CORE-OM)]. *Psihologijske Teme, 23*(2), 265–288.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): A brief measure of Dark Personality Traits. *Assessment, 21*(1), 28–41. https://doi.org/10.1177/1073191113514105

Khan, A. K., Bell, C. M., & Quratulain, S. (2017). The two faces of envy: Perceived opportunity to perform as a moderator of envy manifestation. *Personnel Review, 46*(3), 490–511. https://doi.org/10.1108/PR-12-2014-0279

Kline, R. B. (2016). *Methodology in the social sciences. Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kowalski, C. M., Rogoza, R., Vernon, P. A., & Schermer, J. A. (2018). The Dark Triad and self-presentation variables of socially desirable responding and self-monitoring. *Personality and Individual Differences, 120*, 234–237. https://doi.org/10.1016/j.paid.2017.09.007

Kwiatkowska, M. M., Rogoza, R., & Volkodav, T. (2020). Psychometric properties of the Benign and Malicious Envy Scale: Assessment of structure, reliability, and measurement invariance across the United States, Germany, Russia, and Poland. *Current Psychology*. Advance online publication. https://doi.org/10.1007/s12144-020-00802-4

Lange, J., Blatz, L., & Crusius, J. (2018). Dispositional envy: A conceptual review. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *SAGE Handbook of personality and individual differences* (pp. 424–440). SAGE.

Lange, J., & Crusius, J. (2015). Dispositional envy revisited: Unraveling the motivational dynamics of benign and malicious envy. *Personality and Social Psychology Bulletin, 41*(2), 284–294. https://doi.org/10.1177/0146167214564959

Lange, J., Crusius, J., & Hagemeyer, B. (2016). The evil queen's dilemma: Linking narcissistic admiration and rivalry to benign and malicious envy. *European Journal of Personality, 30*(2), 168–188. https://doi.org/10.1002/per.2047

Lange, J., Paulhus, D. L., & Crusius, J. (2018). Elucidating the dark side of envy: Distinctive links of benign and malicious envy with dark personalities. *Personality and Social Psychology Bulletin, 44*(4), 601–614. https://doi.org/10.1177/0146167217746340

Lange, J., Weidman, A. C., & Crusius, J. (2018). The painful duality of envy: Evidence for an integrative theory and a meta-analysis on the relation of envy and schadenfreude. *Journal of Personality and Social Psychology, 114*(4), 572–598. https://doi.org/10.1037/pspi0000118

Međedović, J., Čolović, P., Dinić, B. M., & Smederevac, S. (2019). The HEXACO Personality Inventory: Validation and psychometric properties in the Serbian language. *Journal of Personality Assessment, 101*(1), 25–31. https://doi.org/10.1080/00223891.2017.1370426

Miceli, M., & Castelfranchi, C. (2007). The envious mind. *Cognition and Emotion, 21*(3), 449–479. https://doi.org/10.1080/02699930600814735

Mosquera, P. M. R., Parrott, W. G., & de Mendoza, A. H. (2010). I fear your envy, I rejoice in your coveting: On the ambivalent experience of being envied by others. *Journal of Personality and Social Psychology, 99*(5), 842–854. https://doi.org/10.1037/a0020965

O'Meara, A., Davies, J., & Hammond, S. (2011). The psychometric properties and utility of the Short Sadistic Impulse Scale (SSIS). *Psychological Assessment, 23*(2), 523–531. https://doi.org/10.1037/a0022400

Parrott, W. G., & Smith, R. H. (1993). Distinguishing the experiences of envy and jealousy. *Journal of Personality and Social Psychology, 64*(6), 906–920. https://doi.org/10.1037/0022-3514.64.6.906

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and Item Response Theory analyses. *Journal of Statistical Software, 17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

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

Sawada, M., & Fujii, T. (2016). Do envious people show better performance? Focusing on the function of benign envy as a personality trait. *Shinrigaku kenkyu: The Japanese Journal of Psychology, 87*(2), 198–204. https://doi.org/10.4992/jjpsy.87.15316

Smith, R. H., & Kim, S. H. (2007). Comprehending envy. *Psychological Bulletin, 133*(1), 46–64. https://doi.org/10.1037/0033-2909.133.1.46

Smith, R. H., Parrott, W. G., Diener, E. F., Hoyle, R. H., & Kim, S. H. (1999). Dispositional envy. *Personality and Social Psychology Bulletin, 25*(8), 1007–1020. https://doi.org/10.1177/01461672992511008

Van de Ven, N., Hoogland, C. E., Smith, R. H., Van Dijk, W. W., Breugelmans, S. M., & Zeelenberg, M. (2015). When envy leads to schadenfreude. *Cognition and Emotion, 29*(6), 1007–1025. https://doi.org/10.1080/02699931.2014.961903

Van de Ven, N., Zeelenberg, M., & Pieters, R. (2009). Leveling up and down: the experiences of benign and malicious envy. *Emotion, 9*(3), 419–429. https://doi.org/10.1037/a0015669

Wilkin, C. L., & Connelly, C. E. (2015). Green with envy and nerves of steel: Moderated mediation between distributive justice and theft. *Personality and Individual Differences, 72*, 160–164. https://doi.org/10.1016/j.paid.2014.08.039

Zizzo, D. J., & Oswald, A. J. (2001). Are people willing to pay to reduce others' incomes? *Annales d'Economie et de Statistique, 63–64*, 39–65. https://doi.org/10.2307/20076295

History
Received September 21, 2020
Revision received January 12, 2021
Accepted January 27, 2021
Published online April 22, 2021
EJPA Section / Category Personality

Acknowledgment
We would like to thank Prof. Jens Lange for thoughtful comments and suggestion for improvements of our paper.

Publication Ethics
For study 1, the Research Ethics Board approval was obtained from the Commission of Ethics and Bioethics at Cardinal Stefan Wyszyński University in Warsaw, Poland (registration number: KEiB – 14/2017).
Study 2 has been approved by the Ethical Committee of the Department of Psychology, Faculty of Philosophy, University of Novi Sad, Serbia.

Open Data
Authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including R code at https://osf.io/3msne/ (Dinić & Branković, 2021).
Authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology (Dinić & Branković, 2021).
The studies were not preregistered.

Funding
Data from Study 1 were collected within the projects of the Polish Minister of Science and Higher Education [Grant No. 0101/DIA/2017/46] and the National Science Centre, Poland [Grant No. 2015/19/N/HS6/00685]. Study 2 was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia [Grant ON179006].

Bojana M. Dinić
Department of Psychology
Faculty of Philosophy
University of Novi Sad
Dr. Zorana Đinđića 2
21000 Novi Sad
Republic of Serbia
bojana.dinic@ff.uns.ac.rs

Multistudy Report

Measurement Invariance of the SOC-13 Sense of Coherence Scale Across Gender and Age Groups

Dennis Grevenstein[1] and Matthias Bluemke[2]

[1]Psychological Institute, Heidelberg University, Germany
[2]GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

Abstract: Sense of coherence (SOC) describes an individual's ability to deal with life challenges (manageability), comprehend the environment (comprehensibility), and perceive life and its challenges as meaningful (meaningfulness). We examine measurement invariance (MI) of the SOC-13 scale across gender and age groups in a matched sample of N = 1,816 (50% females; age range 16–83 years). A two-factor model, with a common factor for manageability/comprehensibility items and a second factor for meaningfulness items, best represented the SOC-13 in all groups. Full metric, partial scalar, and full strict invariance held across gender groups. Across age groups, full metric, partial scalar, and partial strict invariance could be established. We conclude that SOC-13 is a reliable and valid measure. Measurement is comparable across gender and age.

Keywords: sense of coherence, factorial validity, measurement invariance

Sense of coherence (SOC) represents the core concept in Antonovsky's salutogenic theory (Antonovsky, 1979, 1987). Antonovsky proposed that health and disease should not be considered binary on/off states. Instead, he assumed that every human being can be placed on a larger continuum between health and disease, and SOC represents the most crucial concept that helps people move to the health end of the continuum. He argued that stressors are so ubiquitous in life that humans need a range of *general resistance resources* to deal with stressors and life challenges (Antonovsky, 1987). Thus, salutogenesis offers a resource-oriented perspective on health.

SOC comprises three components:

(1) manageability describes an individual's feeling that one has the necessary behavioral capacity and resources (e.g., skills, family, a social network) to deal with life challenges;
(2) comprehensibility is a cognitive aspect that represents an individual's perception that internal aspects and external situations and events are rational and understandable, and that even chaotic situations can be structured; and
(3) meaningfulness reflects that life has some kind of (emotional) meaning, so that its demands and challenges are worthy of investment and engagement (Eriksson & Mittelmark, 2017).

Antonovsky considered SOC an "orientation to life," rather than a temperamental personality trait (Antonovsky, 1987). He theoretically explained the development of SOC as a dynamic process up to age 30. Up to this age, SOC is supposed to be fluctuant, malleable, and shaped by experience. Consistency (enhancing comprehensibility), load-balancing (enhancing manageability), and participation in decision-making (enhancing meaningfulness) are all supposed to foster SOC across the developmental phase (Antonovsky, 1987). This theoretical view was backed up by empirical research according to a recent review of 37 studies from 14 countries, in which the authors concluded that "[t]he … surveyed studies support the conceptualization of the SOC construct as an important personal resource that develops during childhood" (Idan et al., 2017, p. 118). Adolescence is seen as a particularly sensitive phase. Notably, the quality of parent-child relationships (Rivera et al., 2013) and a child-centered parenting style (Feldt et al., 2005) have been shown to be main predictors of SOC. SOC was also correlated with the quality of family relationships in later life (Grevenstein et al., 2019).

The clinical utility of SOC has been shown many times. SOC has been linked to good mental health and health-related behavior (Eriksson & Lindström, 2006), general psychological well-being (Nilsson et al., 2010), depression (Haukkala et al., 2013), anxiety (Moksnes, Espnes, & Haugan, 2013), general psychological distress (Grevenstein, Aguilar-Raab, et al., 2016), burnout (Grevenstein et al., 2018), satisfaction with life (Moksnes, Løhre, & Espnes, 2013), and substance use of tobacco, alcohol, and cannabis (Grevenstein, Bluemke, et al., 2016).

Historically, SOC has been criticized for its alleged similarity to classic personality traits like the Big Five personality factors, specifically neuroticism or emotional stability (Geyer, 1997). High negative correlations with neuroticism have been found, as well as smaller positive correlations with extraversion, agreeableness, and conscientiousness (Feldt, Metsäpelto, et al., 2007; Hochwälder, 2012; Kase et al., 2018). Overall, the Big Five can predict up to 40% of the SOC variance (Hochwälder, 2012). SOC has also shown surprisingly high longitudinal stability, almost comparable to temperamental personality traits. For adults, test-retest reliabilities of .78 over 1 year, .59–.67 over 5 years, and .54 over 10 years emerged (Eriksson & Lindström, 2005). Even in adolescence, SOC was found to be moderately stable. Honkinen and colleagues (2008) found only minor changes of mean SOC scores between ages 15 and 18 longitudinally. SOC at age 15 also predicted SOC scores at age 24 longitudinally at $\beta = .59$ on a latent level (Grevenstein & Bluemke, 2017). Nonetheless, SOC has a unique value for the prediction of health outcomes. SOC has shown incremental validity for the prediction of health-related outcomes above and beyond the Big Five traits (Grevenstein et al., 2018; Grevenstein & Bluemke, 2015), dispositional optimism, resilience, self-compassion (Grevenstein, Aguilar-Raab, et al., 2016), and mindfulness (Grevenstein et al., 2018).

Taken together, current results indicate that changes in SOC cannot be simply traced back to a sensitive period at a specific age. Feldt and colleagues compared two groups of 25- to 29-years-old and 35- to 40-years-old participants regarding their longitudinal change in SOC (Feldt et al., 2003). Both groups improved in a similar fashion with the older group showing very slightly lower SOC. In the much larger Finnish HeSSup study, participants over the age of 30 showed consistently higher mean SOC scores than participants under the age of 30 (Feldt, Lintula, et al., 2007). Silverstein and Heap (2015) showed that mean SOC scores increased continuously with age for older adults beyond the age of 55. Yet all this work presupposed (and left untested) the belief that the SOC scores can legitimately be compared across age groups. It is a rather strong assumption to think that the interpretation of SOC items and the applicability of SOC scale throughout ontogenesis are possible without measurement bias.

The SOC-13 Scale

The most popular measure of SOC is the 13-item SOC scale, originally published by Antonovsky (1987). The scale has been validated in a range of later studies (Antonovsky, 1993) and is widely accepted as a reliable and valid measure of SOC (Eriksson & Lindström, 2005). Still, the factorial validity of the SOC-13 scale has also been debated extensively in the past. Antonovsky developed the scale at a time before the general availability of software packages for structural equation modeling (SEM) and confirmatory factor analysis (CFA). Nonetheless, several studies have demonstrated the theoretically derived three-factor structure of the SOC scale (Feldt et al., 2000, 2003), though difficulties emerged. In many cases, items of the SOC scale had to be dropped or the measurement model had to be modified (Feldt et al., 2000, 2003).

One modification to the factor model has consistently and substantially improved model fit (Feldt et al., 2003; Moksnes & Haugan, 2014; Richardson et al., 2007): A residual correlation between items #2 and #3 (#2 "… were you surprised by the behavior of people who you thought you knew well?"; #3: "… have people you counted on disappointed you?") has been interpreted to reflect an additional aspect of interpersonal trust shared between the items (Frenz et al., 1993). Only recently have several studies replicated the intended three-factor model with all items included while allowing for one pair of correlated residuals between items 2 and 3 (Grevenstein, Aguilar-Raab, et al., 2016; Moksnes & Haugan, 2014; Stern et al., 2019). In all cases the comprehensibility and manageability factors correlated so highly ($r > .94$) that one can question if they are truly conceptually distinct.

As an alternative, a more parsimonious two-factor model has been proposed with one factor spanning comprehensibility and manageability items, and with meaningfulness constituting the second factor, though not to be mistaken as a secondary factor of lesser importance (Grevenstein, Aguilar-Raab, et al., 2016; Grevenstein & Bluemke, 2017; Zimprich et al., 2006). Despite its parsimony, this two-factor model has also shown superior predictive validity (Grevenstein, Aguilar-Raab, et al., 2016). One explanation that suggests itself is that comprehensibility and manageability are reciprocal aspects of conquering life stressors, yet meaningfulness constitutes a distinct, but an equally important component that provides the motivation to mobilize any coping resources (Antonovsky, 1987). This is in line with research showing that having a feeling of purpose in life helps people rise above mental health issues (Park, 2010).

Measurement Invariance

A comparison of SOC scores across situational contexts, measurement times, or groups of participants requires that the measurement model is valid across the different subsamples (Vandenberg & Lance, 2000). One needs to ascertain that manifest SOC scores reflect the same latent construct to the same degree. Stability indices and comparisons of manifest mean scores are only valid across different groups as far as strict measurement invariance (MI) can be established (Chen, 2008). Observed differences in scale means have to reflect true differences in latent means, not different item utilization or item difficulties. Otherwise, latent group differences or associations between a latent variable and external criteria might be explained by dissimilar measurement.

MI of the SOC scale has been addressed in the past, with some limitations. Comparing samples of Caucasian Americans and Asian Americans, Stein and colleagues (2006) had to drop items to achieve an acceptable model fit. Across two age groups of adolescents (12–14 years and 15–18 years), Zimprich et al. (2006) showed strict MI of a two-factor model, yet they retained all the items of the SOC-13 scale. Feldt and colleagues showed longitudinal MI across a 5-year span but tested only the invariance of factor loadings and item residuals (Feldt, Lintula, et al., 2007). Grevenstein and Bluemke (2017) showed longitudinal MI at age 15 and age 24, with partial scalar and strict invariance for a two-factor model that included all items. Luyckx and colleagues (2012), again, dropped two items from the SOC-13 scale when investigating MI across age and gender. Results supported scalar MI across ages 14–30 years. The authors also declared scalar invariance across gender, yet described a drop in model fit that – when following common heuristics for MI strictly – exceeded accepted cut-offs. Unfortunately, no detailed analyses of partial MI were provided, so model misfit could not be attributed to specific items. Hittner (2007) tested MI across gender when applying a single-factor model. Though all items were retained, only configural and metric MI were established, and scalar or strict MI were not even tested. To summarize, the field seems to be in a state of disarray, hence the need for a systematic investigation of MI of the SOC-13 scale across age and gender with the modified two-factor model that has consistently emerged in previous work.

Methods

Procedure and Participants

We pooled several samples collected between 2014 and 2016 to analyze the data presented in this study. In all studies, data were collected in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Though studies carried out at the Psychological Institute of the University of Heidelberg were granted an exemption from having to be run past the ethical review board at the time, unless research grant proposals were about to be submitted or sensitive topics were involved, informed consent had been obtained from all participants, and participation was completely voluntary. Most participants had completed short online studies without compensation.

We aimed to select a suitable sample for analysis from a total pool of N = 4,154 (78.1% female) German participants spanning a wide age range of 13–83 years. We used propensity score matching and the *R*-package "MatchIt" with its "nearest" algorithm (Ho et al., 2007) to select a sample with balanced gender groups matched for age. We initially had to exclude 13 participants from the pool, because they had not disclosed their gender. The final sample included N = 1,816 participants (50.0% female). Sample characteristics are displayed in Table 1. As intended, men and women did not differ with regard to their age, $t(1{,}814) = 0.09$, $p = .84$. Based on theoretical grounds we divided the sample into three age groups: Youth and young adults (age = 16–29 years; n = 1,008); adults (age = 30–49 years; n = 484), and older adults including seniors (age > 50 years; n = 324). There were no missing SOC-13 values, mostly due to the fact that the online participants were technically required to provide answers to every item; otherwise, they were considered drop-out participants, because we had assured them that refusing to participate (any further) was possible at any time.

Measures

We used the 13-item version of Antonovsky's original Orientation to Life scale (Antonovsky, 1987). The German adaptation was provided by Schumacher and colleagues (2000). The scale includes five comprehensibility items (e.g., "Has it happened in the past that you were surprised by the behavior of people whom you thought you knew well?"), four manageability items (e.g., "Has it happened that people whom you counted on disappointed you?"), and four meaningfulness items (e.g., "Do you have the feeling that you do not really care about what goes on around you?"). Answers were provided on 7-point rating scales, marked from 1 = *very often* to 7 = *very seldom or never* most of the time. Items #1, #2, #3, and #7 were recoded before computing mean scores.

Statistical Analysis

We used SPSS 22 for descriptives and Mplus 7.4 (Muthén & Muthén, 1998–2012) for the CFAs. The arbitrariness of

Table 1. Descriptives, reliability, and standardized factor loadings for accepted models in the total sample (CFA), and subgroups gender (MGCFA 5) and age (MGCFA 5a)

	Total	Female	Male	Age 16–29	Age 30–49	Age 50–83
	N = 1,816	*n* = 908	*n* = 908	*n* = 1,008	*n* = 484	*n* = 324
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Age	33.48 (14.29)	33.45 (14.22)	33.51 (14.38)	22.83 (2.98)	38.87 (5.95)	58.55 (6.30)
SOC mean	4.75 (0.98)	4.73 (1.00)	4.77 (0.96)	4.51 (0.95)	4.87 (0.96)	5.32 (0.82)
SOC1	5.51 (1.45)	5.77 (1.35)	5.24 (1.51)	5.22 (1.52)	5.73 (1.33)	6.09 (1.18)
SOC2	4.02 (1.59)	3.88 (1.59)	4.15 (1.58)	4.07 (1.60)	3.95 (1.57)	3.94 (1.59)
SOC3	3.90 (1.69)	3.67 (1.67)	4.13 (1.68)	3.86 (1.73)	3.87 (1.70)	4.06 (1.54)
SOC4	5.25 (1.37)	5.40 (1.31)	5.11 (1.42)	5.16 (1.38)	5.17 (1.42)	5.68 (1.19)
SOC5	4.82 (1.62)	4.83 (1.61)	4.82 (1.64)	4.64 (1.62)	4.83 (1.67)	5.38 (1.45)
SOC6	4.94 (1.56)	4.96 (1.53)	4.93 (1.59)	4.57 (1.57)	5.17 (1.53)	5.76 (1.16)
SOC7	5.05 (1.25)	5.11 (1.23)	5.00 (1.26)	4.82 (1.22)	5.18 (1.25)	5.58 (1.13)
SOC8	4.67 (1.71)	4.53 (1.75)	4.81 (1.65)	4.18 (1.69)	5.05 (1.59)	5.65 (1.34)
SOC9	4.38 (1.82)	4.22 (1.86)	4.54 (1.76)	4.03 (1.80)	4.56 (1.81)	5.18 (1.58)
SOC10	4.43 (1.68)	4.36 (1.71)	4.50 (1.65)	4.25 (1.68)	4.40 (1.67)	5.03 (1.57)
SOC11	4.69 (1.48)	4.69 (1.52)	4.69 (1.44)	4.44 (1.51)	4.87 (1.43)	5.23 (1.28)
SOC12	4.91 (1.68)	5.03 (1.65)	4.78 (1.71)	4.52 (1.68)	5.15 (1.65)	5.74 (1.35)
SOC13	5.20 (1.60)	5.09 (1.67)	5.31 (1.52)	4.89 (1.67)	5.42 (1.49)	5.82 (1.26)
α	.87	.87	.86	.85	.87	.85
	λ	λ	λ	λ	λ	λ
SOC1	.35	.35	.35	.29	.34	.31
SOC2	.33	.33	.33	.36	.36	.33
SOC3	.47	.48	.48	.49	.49	.46
SOC4	.56	.56	.56	.55	.55	.50
SOC5	.57	.57	.57	.56	.56	.52
SOC6	.66	.65	.65	.61	.61	.71
SOC7	.73	.72	.72	.72	.72	.67
SOC8	.75	.75	.75	.71	.71	.76
SOC9	.77	.77	.77	.75	.75	.72
SOC10	.73	.73	.73	.74	.74	.70
SOC11	.49	.49	.49	.47	.47	.44
SOC12	.80	.80	.80	.79	.79	.75
SOC13	.69	.69	.69	.66	.66	.70
CR	.89	.89	.89	.88	.88	.87

Note. CR = Raykov's composite reliability.

using cut-offs notwithstanding, we evaluated model fit by (1) the – ideally nonsignificant – χ^2-test (Bentler & Bonett, 1980); (2) the comparative fit index (CFI) with values of .90/.95 and above indicating appropriate/good model fit (Bentler, 1990; Hu & Bentler, 1999); (3) the root mean square error of approximation (RMSEA) with values of .00–.05/.06–.08/.09–.10 indicating excellent/adequate/poor model fit (Browne & Cudeck, 1993); and (4) the standardized root mean square residual (SRMR) with values less than .08 (Hu & Bentler, 1999) or .05 (Schumacker & Lomax, 2010) considered to reflect good or excellent fit.

When comparing different models based on the same data and variables, we prefer the Akaike information criterion (AIC) and the Bayesian Information Criterion (BIC), which can be used to compare the quality of different models. Lower scores indicate better model fit (Akaike, 1987), and differences larger than 10 indicate "very strong" differences (Raftery, 1995). AIC commonly emphasizes accuracy, whereas BIC provides the best trade-off between accuracy and parsimony, which is most relevant for MI testing procedures. In line with prior research, we used Robust Maximum Likelihood (MLR) for parameter estimation.

For an estimation of reliability, we provide – for descriptive purposes – Cronbach's α, but also Raykov's composite reliability (CR; Raykov, 1997) as an SEM-based reliability estimate. As the SOC-13 scale lacks essential tau-equivalence and strict unidimensionality, Cronbach's α will be biased, whereas composite reliability is unbiased and preferable (Graham, 2006).

MI can be tested via a series of nested, increasingly restricted confirmatory factor-analytical (CFA) models (Meredith, 1993; Vandenberg & Lance, 2000). MI across independent groups, such as age or gender, is investigated using multiple-groups CFA (Brown, 2006). If age trends were assessed within the same group of participants, a longitudinal design for testing MI is required (Marsh & Grayson, 1994; Millsap & Cham, 2012). Four increasingly restrictive forms of MI can be tested (Vandenberg & Lance, 2000): (1) Configural MI indicates equal construct dimensionality and item-to-factor patterns across groups. Factor loadings, item intercepts, and residuals may differ. (2) Metric MI requires all factor loadings to be equal across groups. (3) Scalar MI additionally constrains all item intercepts to be equal across groups. (4) Strict MI further assumes equal residual variances. If at least some levels of MI can be established, it is possible to further investigate the invariance of structural parameters. Assuming metric invariance held, we wanted to test (5) the invariance of factor variances and covariances, including the residual correlation between items #2 and #3. As the last step, (6) the equality of factor means can be tested, assuming scalar invariance held.

Different levels of MI have ramifications for the applicability of scales and the comparability of scores. Metric MI indicates that latent variables representing the substantive factor reflect the same psychological meaning. Technically, metric MI implies that item scores are based on the same unit of measurement, which allows for a comparison of (latent) variance/covariance structures. Scalar MI denotes that item difficulties are comparable, which allows for a comparison of (latent) means. Strict MI indicates an equal impact of sources of item specificity (e.g., unreliability). When strict MI holds, differences in manifest variables are due to true differences in the latent variables, rather than item-specific measurement error. If strict MI holds, direct comparisons of manifest scale means are possible. In combination with equal latent variances, strict MI also implies that measurement reliability (proportion of true score variance to total variance) is comparable.

Tests of MI have often shown that invariance levels beyond metric invariance are hard to achieve (Schmitt & Kuljanin, 2008), but even lower levels of invariance may support comparable measurement (Tran, 2009). Traditionally, partial MI can be investigated if some item parameters (either loadings or intercepts) are non-invariant. For example, partial metric MI does not require all, but two, of the factor loadings to be equal (one anchor item's plus another invariant item's loadings). Partial scalar MI (one anchor item's plus another invariant item's intercepts) is statistically sufficient to compare latent means (Byrne et al., 1989; Lubke & Dolan, 2003). In a review on MI testing, Schmitt and Kuljanin (2008) concluded that "partial invariance made little difference in the estimates of structural model parameters." Candidate items that supply invariant model parameters can be found on the basis of χ^2-based Modification Indices (ModInd; Byrne et al., 1989). Researchers are advised to relax parameters one at a time to see if model fit can be improved and if partial MI can be established. A ModInd around 3.84 ($df = 1$) is just statistically significant at an arbitrary level of $p = .05$ for Type-I errors. Researchers are advised to look for modifications that substantially exceed this threshold (Brown, 2006). We generally followed these recommendations, and all modifications were executed based on ModInd. ModInd were used when a specific invariance test failed, but partial invariance was still an option. In an iterative manner, the largest ModInd (typically >10) was used to identify the model parameter most in need of being freed from a cross-group equivalence constraint. The partial invariance model was then inspected for acceptable model fit.

In MI testing, the alternative models which are nested in less constrained baseline models are compared based on χ^2-difference tests. MLR uses scaled χ^2-scores, but χ^2-difference scores are not χ^2-distributed themselves, necessitating Satorra-Bentler scaled χ^2-difference tests (Satorra, 2000; Satorra & Bentler, 2001). Any χ^2-tests or χ^2-difference tests are likely to reach significance due to our large sample. Independent from the sample size, model fit indices can be used to evaluate MI analyses. Going from one step to the next, a drop in CFI less or equal to .010 is conventionally considered acceptable unless there is a concurrent increase of RMSEA greater than +.015 (Chen, 2007; Cheung & Rensvold, 2002). However, strictly adhering to cut-offs when examining ΔCFI and ΔRMSEA is problematic when the simulation parameters used for deriving the cut-offs somehow differ from the present case (Fan & Sivo, 2009; Saris et al., 2009). A better alternative is to look for lower BIC values that indicate a better tradeoff between accuracy and parsimony, irrespective of sample size and model complexity.

Results

Descriptive Data Analysis

Men and women hardly differed at all regarding their mean SOC scores, $t(1,814) = 0.765$, $p = .45$, $d = 0.04$. Across age groups, SOC scores increased almost linearly with age, $F(2, 1,813) = 98.41$, $p < .001$, $\eta^2_p = .10$. With regard to reliability estimates, Cronbach's αs were consistently high in all groups. For comparison with a representative German sample, we computed a mean score of SOC means for 808 participants in the two older age groups ($M_{age} = 46.76$). SOC

means were comparable to the reference sample (N = 2,005, M_{age} = 50.03) reported by Schumacher and colleagues (2000): M = 5.05 (SD = 0.93) versus M_{ref} = 5.01 (SD_{ref} = 0.89), d = 0.04.

Confirmatory Factor Analyses and Measurement Invariance

Matching previous researchers' considerations, we first compared a three-factor model and a two-factor model (across the whole sample), assuming a pair of correlated residuals between items #2 and #3. The three-factor model fitted the data well, $\chi^2(61) = 376.22$, $p < .001$, RMSEA = .053, CI_{90} = [.048, .059], CFI = .952, SRMR = .036, but so did the two factor-model, $\chi^2(63) = 383.77$, $p < .001$, RMSEA = .053, CI_{90} [.048, .058], CFI = .951, SRMR = .036. The factors for manageability and comprehensibility were statistically nearly indistinguishable (r = .96). Having replicated the recently emerged standard model for the SOC-13 scale, we accepted the two-factor model as the basis for the following MI analyses.

Results and model fit indices of the MGCFAs are depicted in Table 2. Testing first the invariance across genders, the initial configural invariance model showed a good model fit. Constraining factor loadings to be equal (metric MI) did not impair model fit. We observed a noticeable drop in model fit when testing scalar MI, so we investigated modification indices (ModInd) to check for potential adjustments of the model. The intercept of item #1 (ModInd = 37.80, $\Delta\chi^2$ = 38.47) "Do you have feeling that you do not really care about what goes on around you?" was not invariant across genders. After relaxing its equality constraint, partial scalar MI held. Within the context of a partial scalar invariant model, we next examined strict invariance by constraining residuals to be equal across groups. The decrease in model fit was well within acceptable levels. At the level of structural parameters, enforcing equal variances and covariances also did not harm model fit. At last, we tested for equal factor means. The drop in CFI was still below .010 and RMSEA increased only slightly. Still, ModInd indicated unequal factor means for both factors (meaningfulness: ModInd: 43.69; comprehensibility/manageability: ModInd: 41.34). On the basis of the variance invariance model (M5), we can quantify the estimated latent mean differences. Compared to females, males had lower (unstandardized) latent means on meaningfulness (Est. = −0.193, SE = 0.054, $p < .001$, d = .12) and higher latent means on comprehensibility/manageability (Est. = 0.161, SE = 0.051, p = .002, d = .10).

We then investigated MI across age groups. Again, the configural MI model showed good model fit. Metric invariance also held unconditionally. The scalar MI model showed a substantial decrease in model fit. ModInd indicated unequal intercepts for several items among the young adults: item #8 (ModInd = 49.29, $\Delta\chi^2$ = 49.71) "Do you have very mixed-up feelings and ideas?", item #6 (ModInd = 33.36, $\Delta\chi^2$ = 33.58) "Do you have the feeling that you are in an unfamiliar situation and do not know what to do?", item #1 (ModInd = 31.63, $\Delta\chi^2$ = 31.10), item #4 (ModInd = 16.98, $\Delta\chi^2$ = 8.41) "Until now your life has had … clear goals", item #10 (ModInd = 16.02, $\Delta\chi^2$ = 16.85) "Many people – even those with a strong character – sometimes feel like sad sacks (losers) in certain situations. How often have you felt this way in the past?", item #2 (ModInd = 11.17, $\Delta\chi^2$ = 11.98) "Has it happened in the past that you were surprised by the behavior of people whom you thought you knew well?", and item #3 (ModInd = 25.17, $\Delta\chi^2$ = 26.09) "Has it happened that people whom you counted on disappointed you?". After these modifications, partial scalar MI could be established. Notably, all modifications pertained to the young adult group. We next tested strict MI by constraining all residuals to be equal across age groups, yet once more model fit dropped below accepted cut-offs. In the senior age group, the residuals of item #6 (ModInd = 31.81, $\Delta\chi^2$ = 42.93), item #8 (ModInd = 17.92, $\Delta\chi^2$ = 20.67), and item #13 (ModInd = 14.63, $\Delta\chi^2$ = 15.02) were non-invariant, as was item #1 in the young adult group (ModInd = 25.85, $\Delta\chi^2$ = 26.28). After modifications partial strict MI could be established.

At the level of structural parameters, we constrained all variances and covariances to be equal across groups. Model fit dropped slightly. Most notably, SRMR increased by .025. ModInd indicated unequal variances for both SOC factors in the senior group: meaningfulness (ModInd = 16.82, $\Delta\chi^2$ = 169.42) and comprehensibility/manageability (ModInd = 2,009.67, $\Delta\chi^2$ = 1029.24). Finally, we constrained latent means to be equal. As expected, model fit clearly decreased. On the basis of the accepted partial variance invariance model, we could estimate latent mean differences. Compared to the young adult group, the adult age group had higher (unstandardized) scores on latent meaningfulness (Est. = 0.297, SE = 0.071, $p < .001$, d = .19) and comprehensibility/manageability (Est. = 0.410, SE = 0.065, $p < .001$, d = .29) factors. Differences between the young and senior groups were even stronger for meaningfulness (Est. = 0.799, SE = 0.075, $p < .001$, d = .59) and comprehensibility/manageability (Est. = 0.902, SE = 0.068, $p < .001$, d = .74).

Discussion

The present research investigated measurement invariance (MI) of SOC as measured by the SOC-13 scale across age and gender. In line with prior research, a three-factor model

Table 2. Measurement invariance of SOC across gender and age

MGCFA comparison	Equal loadings	Equal intercepts	Equal residuals	Equal variances	Equal means	*df*	χ^2	Δdf	$\Delta\chi^2$	CFI	RMSEA [CI_{90}]	SRMR	AIC	BIC
Gender														
1. Configural invariance						126	432.58**	–	–	.954	.052 [.046, .057]	.037	80,385	80,837
2. Metric invariance	×					137	464.51**	11	31.27**	.951	.051 [.046, .056]	.045	80,396	80,787
3. Scalar invariance	×	×				148	560.16**	11	104.82**	.938	.055 [.051, .060]	.053	80,480	80,810
3a. Partial scalar invariance	×	×				147	521.69**	10	61.04**	.944	.053 [.048, .058]	.051	80,438	80,773
4. Strict invariance	×	×	×			160	563.62**	13	41.78**	.939	.053 [.048, .057]	.059	80,458	80,722
5. Structural: (co)variances	×	×	×	×		**164**	**575.09****	**4**	**11.44***	**.938**	**.053 [.048, .057]**	**.063**	**80,463**	**80,706**
6. Structural: means	×	×	×	×	×	166	629.58**	2	62.77**	.930	.055 [.051, .060]	.072	80,521	80,753
Age														
1. Configural invariance						189	530.97**	–	–	.944	.055 [.049, .060]	.042	79,969	80,646
2. Metric invariance	×					211	565.13**	22	35.37*	.942	.053 [.044, .058]	.052	79,968	80,524
3. Scalar invariance	×	×				233	787.46**	22	239.97**	.909	.058 [.058, .068]	.066	80,176	80,611
3a. Partial scalar invariance	×	×				226	609.73**	15	45.21**	.937	.053 [.048, .058]	.055	79,985	80,458
4. Strict invariance	×	×	×			252	781.58**	26	159.45**	.913	.059 [.054, .064]	.100	80,147	80,478
4a. Partial strict invariance	×	×	×			248	676.69**	22	66.47**	.930	.053 [.049, .058]	.070	86,294	86,599
5. Structural: (co)variances	×	×	×	×		256	709.98**	8	31.34**	.925	.054 [.049, .059]	.095	80,052	80,361
5a. Partial (co)variances	×	×	×	×		**254**	**683.99****	**6**	**8.17**	**.929**	**.053 [.048, .058]**	**.078**	**80,023**	**80,342**
6. Structural: means	×	×	×	×	×	258	843.24**	4	160.87**	.904	.061 [.057, .066]	.136	80,201	80,499

Note. Model fit: *$p < .05$; **$p < .001$. Accepted models printed in bold. χ^2 values are Satorra-Bentler scaled.

fitted the data, yet resulted in a very high correlation between the manageability and comprehensibility factors. For reasons of parsimony, we tested MI on the basis of a two-factor model (Grevenstein, Aguilar-Raab, et al., 2016; Grevenstein & Bluemke, 2017; Zimprich et al., 2006). The alleged correlated residuals between items #2 and #3 replicated across all subgroups (Feldt et al., 2003; Grevenstein, Aguilar-Raab, et al., 2016; Moksnes, Espnes, et al., 2013).

Across gender groups, full metric MI, partial scalar MI, and strict MI could be established. Even at the level of structural parameters, the invariance of all variances and covariances could be shown. Taken together, only one intercept (item #1) differed, and females endorsed the item more readily. Only minor differences in latent means emerged. Even manifest scale means appear to be quite comparable across genders without introducing a large amount of bias.

Across age groups, full metric MI, partial scalar MI, and partial strict MI held. Non-invariance at the scalar level was due to items in the young adult group, with seven items having non-invariant intercepts. This finding is unprecedented, but highly relevant, as not a single study has investigated SOC's MI across this wide age range before. This finding is highly illuminating for salutogenic theory. Even though non-invariance may be due to uniform item bias, or disparate use of items by young cohort members, a more theoretical explanation pertains to the development of SOC at young adult age.

Antonovsky described the development of SOC as a dynamic process up to the age of 30 (Antonovsky, 1987). He assumed SOC to be fully developed in later years. It has long been accepted that people face different developmental tasks across various age stages (Havighurst, 1972). Full metric MI for the SOC-13 provided the first evidence that the same psychological construct is measured irrespective of the age of participants. And yet, the different components underlying SOC (at the factor and item level) are unevenly important at various age stages. One can easily fathom that most life challenges are different for a 20-year-old who just started life on their own from a 50-year-old who has dealt with these challenges and may then be tightly embedded in family structures or professional settings. This logic is in line with recent research on life goals. Much like the challenges that life poses, life goals change across the life-span. The importance of personal-growth, status, and work goals decreased, whereas prosocial-engagement increased in importance (Bühler et al., 2019). Moreover, goal-adjustment, that is, adaptive disengagement and re-engagement capacities were found to predict individuals' well-being and health (Barlow et al., 2019). Our results of partial scalar MI have to be seen through the lens of developmental tasks that influence how people interpret the SOC-13 items from within their age-dependent context, which systematically affects the item difficulty when choosing one of the response options of the non-invariant items. This age-dependent differential item functioning needs to be kept in mind when comparing SOC scores, before arriving at firm conclusions on true differences between groups, or changes across the life-span, that involve young adult age.

Unfortunately, neither could we establish strict MI unconditionally, though only a minority of item residuals had to be estimated freely. Covariances between factors and between the residuals of items #2 and #3 were found to be equal, yet factor variances in the senior group were reduced. The latent means of both meaningfulness and comprehensibility/manageability showed a substantial, and nearly linear, increase along with age. This is in line with prior research, where older participants reported higher SOC means (Feldt, Lintula, et al., 2007; Silverstein & Heap, 2015). This increase of SOC scores refutes Antonovsky's age stability hypothesis, which entails variability and development of generalized resistance resources up to the age of 30, but stability afterward. This finding can easily be reconciled with previous research demonstrating relatively high stability of participants' rank-order. The increase in SOC that may occur in later life affects all people to a similar extent. Consequently, in the future researchers need to be clear about whether they refer to absolute SOC levels (which progress even at higher age) or relative individual differences (which appear quite stable). If one reads Antonovsky to refer to the former, the theoretical supposition has to be rejected, whereas it may be compatible with the latter notion.

Regarding the measurement of SOC, our results shine a positive light on the factorial validity of the SOC-13 scale. For both types of group comparisons, we could establish (partial) strict MI. Most researchers consider partial scalar MI as "good enough" for an interpretation of mean structures (Byrne et al., 1989; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998). Strict MI is even more difficult to achieve, yet the consequences of unequal residuals may be low in practical terms (Lubke & Dolan, 2003; Schmitt & Kuljanin, 2008).

Antonovsky conceived SOC as a meta-construct. As a medical-sociologist, he envisioned numerous potential external factors that could influence SOC (Antonovsky, 1987). Antonovsky assumed that a person's SOC develops in line with a broad range of concepts, including socioeconomic status, social environment, general intelligence, and personal experiences in interpersonal relationships. Our results support the conclusion that despite the complexity of the SOC construct, its measurement across age and gender is – almost surprisingly – comparable.

Limitations

Our study offers progress beyond prior research, as we were able to analyze a large sample with a wide range of age represented in it. Due to the use of propensity score matching, we can confidently assume that the effects of participants' age and gender were not confounded. Still, our sample does not represent a real probability-based sample. Also, it needs to be seen if good results can be achieved in other languages as well.

Conclusions

We confirmed partial strict MI of the SOC-13 scale across gender and age groups, allowing for comparable measurement of SOC. Our results lend support to previous and future comparisons of variance/covariance structures (correlations) and mean structures across groups when based on the SOC-13 scale.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317–332. https://doi.org/10.1007/BF02294359

Antonovsky, A. (1979). *Health, stress, and coping*. Jossey-Bass.

Antonovsky, A. (1987). *Unraveling the mystery of health: How people manage stress and stay well*. Jossey-Bass.

Antonovsky, A. (1993). The structure and properties of the Sense of Coherence scale. *Social Science & Medicine, 36*(6), 725–733. https://doi.org/10.1016/0277-9536(93)90033-Z

Barlow, M. A., Wrosch, C., & McGrath, J. J. (2020). Goal adjustment capacities and quality of life: A meta-analytic review. *Journal of Personality, 88*(2), 307–323. https://doi.org/10.1111/jopy.12492

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Bühler, J. L., Weidmann, R., Nikitin, J., & Grob, A. (2019). A closer look at life goals across adulthood: Applying a developmental perspective to content, dynamics, and outcomes of goal importance and goal attainability. *European Journal of Personality, 33*(3), 359–384. https://doi.org/10.1002/per.2194

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Eriksson, M., & Lindström, B. (2005). Validity of Antonovsky's Sense Of Coherence scale: A systematic review. *Journal of Epidemiology and Community Health, 59*(6), 460–466. https://doi.org/10.1136/jech.2003.018085

Eriksson, M., & Lindström, B. (2006). Antonovsky's Sense of Coherence Scale and the relation with health: A systematic review. *Journal of Epidemiology and Community Health, 60*(5), 376–381. https://doi.org/10.1136/jech.2005.041616

Eriksson, M., & Mittelmark, M. B. (2017). The sense of coherence and its measurement. In M. B. Mittelmark, S. Sagy, M. Eriksson, G. F. Bauer, J. M. Pelikan, B. Lindström, & G. A. Espnes (Eds.), *The handbook of salutogenesis* (pp. 97–106). Springer International Publishing.

Fan, X., & Sivo, S. A. (2009). Using Δgoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling, 16*(1), 54–69. https://doi.org/10.1080/10705510802561311

Feldt, T., Kokko, K., Kinnunen, U., & Pulkkinen, L. (2005). The role of family background, school success, and career orientation in the development of sense of coherence. *European Psychologist, 10*(4), 298–308. https://doi.org/10.1027/1016-9040.10.4.298

Feldt, T., Leskinen, E., Kinnunen, U., & Mauno, S. (2000). Longitudinal factor analysis models in the assessment of the stability of sense of coherence. *Personality and Individual Differences, 28*(2), 239–257. https://doi.org/10.1016/s0191-8869(99)00094-x

Feldt, T., Leskinen, E., Kinnunen, U., & Ruoppila, I. (2003). The stability of sense of coherence: Comparing two age groups in a 5-year follow-up study. *Personality and Individual Differences, 35*(5), 1151–1165. https://doi.org/10.1016/s0191-8869(02)00325-2

Feldt, T., Lintula, H., Suominen, S., Koskenvuo, M., Vahtera, J., & Kivimäki, M. (2007). Structural validity and temporal stability of the 13-item Sense of Coherence Scale: Prospective evidence from the population-based HeSSup study. *Quality of Life Research, 16*(3), 483–493. https://doi.org/10.1007/s11136-006-9130-z

Feldt, T., Metsäpelto, R., Kinnunen, U., & Pulkkinen, L. (2007). Sense of coherence and five-factor approach to personality: Conceptual relationships. *European Psychologist, 12*(3), 165–172. https://doi.org/10.1027/1016-9040.12.3.165

Frenz, A. W., Carey, M. P., & Jorgensen, R. S. (1993). Psychometric evaluation of Antonovsky's Sense of Coherence Scale. *Psychological Assessment, 5*(2), 145–153. https://doi.org/10.1037/1040-3590.5.2.145

Geyer, S. (1997). Some conceptual considerations on the sense of coherence. *Social Science & Medicine, 44*(12), 1771–1779. https://doi.org/10.1016/S0277-9536(96)00286-9

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930–944. https://doi.org/10.1177/0013164406288165

Grevenstein, D., Aguilar-Raab, C., & Bluemke, M. (2018). Mindful and resilient? Incremental validity of sense of coherence over mindfulness and Big Five personality factors for quality of life outcomes. *Journal of Happiness Studies, 19*, 1883–1902. https://doi.org/10.1007/s10902-017-9901-y

Grevenstein, D., Aguilar-Raab, C., Schweitzer, J., & Bluemke, M. (2016). Through the tunnel, to the light: Why sense of coherence covers and exceeds resilience, optimism, and self-compassion.

Personality and Individual Differences, 98, 208–217. https://doi.org/10.1016/j.paid.2016.04.001

Grevenstein, D., & Bluemke, M. (2015). Can the Big Five explain the criterion validity of sense of coherence for mental health, life satisfaction, and personal distress? *Personality and Individual Differences, 77*, 106–111. https://doi.org/10.1016/j.paid.2014.12.053

Grevenstein, D., & Bluemke, M. (2017). Longitudinal factor analysis and measurement invariance of sense of coherence and general self-efficacy in adolescence. *European Journal of Psychological Assessment, 33*, 377–387. https://doi.org/10.1027/1015-5759/a000294

Grevenstein, D., Bluemke, M., & Kroeninger-Jungaberle, H. (2016). Incremental validity of sense of coherence, neuroticism, extraversion, and general self-efficacy: Longitudinal prediction of substance use frequency and mental health. *Health and Quality of Life Outcomes, 14*, Article 9. https://doi.org/10.1186/s12955-016-0412-z

Grevenstein, D., Bluemke, M., Schweitzer, J., & Aguilar-Raab, C. (2019). Better family relationships – higher well-being: The connection between relationship quality and health related resources. *Mental Health & Prevention, 14*, Article 200160. https://doi.org/10.1016/j.mph.2019.200160

Grevenstein, D. (2021). *SOC MI materials. European Journal of Psychological Assessment*. Zenodo. https://doi.org/105281/zenodo.4744808

Haukkala, A., Konttinen, H., Lehto, E., Uutela, A., Kawachi, I., & Laatikainen, T. (2013). Sense of coherence, depressive symptoms, cardiovascular diseases, and all-cause mortality. *Psychosomatic Medicine, 75*(4), 429–435. https://doi.org/10.1097/PSY.0b013e31828c3fa4

Havighurst, R. J. (1972). *Developmental tasks and education*. McKay.

Hittner, J. B. (2007). Factorial invariance of the 13-item Sense of Coherence Scale across gender. *Journal of Health Psychology, 12*(2), 273–280. https://doi.org/10.1177/1359105307074256

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Hochwälder, J. (2012). The contribution of the Big Five personality factors to sense of coherence. *Personality and Individual Differences, 53*(5), 591–596. https://doi.org/10.1016/j.paid.2012.05.008

Honkinen, P., Suominen, S., Helenius, H., Aromaa, M., Rautava, P., Sourander, A., & Sillanpää, M. (2008). Stability of the sense of coherence in adolescence. *International Journal of Adolescent Medicine and Health, 20*(1), 85–91. https://doi.org/10.1515/ijamh.2008.20.1.85

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Idan, O., Braun-Lewensohn, O., Lindström, B., & Margalit, M. (2017). Salutogenesis: Sense of coherence in childhood and in families. In M. B. Mittelmark, S. Sagy, M. Eriksson, G. F. Bauer, J. M. Pelikan, B. Lindström, & G. A. Espnes (Eds.), *The Handbook of Salutogenesis* (pp. 107–121). Springer International Publishing.

Kase, T., Ueno, Y., & Oishi, K. (2018). The overlap of sense of coherence and the Big Five personality traits: A confirmatory study. *Health Psychology Open, 5*(2), 2055102918810654. https://doi.org/10.1177/2055102918810654

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling, 10*(2), 175–192. https://doi.org/10.1207/S15328007SEM1002_1

Luyckx, K., Goossens, E., Apers, S., Rassart, J., Klimstra, T., Dezutter, J., & Moons, P. (2012). The 13-item Sense of Coherence Scale in Dutch-speaking adolescents and young adults: structural validity, age trends, and chronic disease. *Psychologica Belgica, 52*(4), 351–368. https://doi.org/10.5334/pb-52-4-368

Marsh, H. W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling: A Multidisciplinary Journal, 1*(2), 116–145. https://doi.org/10.1080/10705519409539968

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109–127). Guilford Press.

Moksnes, U. K., Espnes, G. A., & Haugan, G. (2013). Stress, sense of coherence and emotional symptoms in adolescents. *Psychology & Health, 29*(1), 32–49. https://doi.org/10.1080/08870446.2013.822868

Moksnes, U. K., & Haugan, G. (2014). Validation of the orientation to life questionnaire in norwegian adolescents, construct validity across samples. *Social Indicators Research, 119*(2), 1105–1120. https://doi.org/10.1007/s11205-013-0536-z

Moksnes, U. K., Løhre, A., & Espnes, G. A. (2013). The association between sense of coherence and life satisfaction in adolescents. *Quality of Life Research, 22*(6), 1331–1338. https://doi.org/10.1007/s11136-012-0249-9

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide: Statistical analysis with latent variables* (7th ed.). Muthén & Muthén.

Nilsson, K. W., Leppert, J., Simonsson, B., & Starrin, B. (2010). Sense of coherence and psychological well-being: Improvement with age. *Journal of Epidemiology and Community Health, 64*(4), 347–352. https://doi.org/10.1136/jech.2008.081174

Park, C. L. (2010). Making sense of the meaning literature: An integrative review of meaning making and its effects on adjustment to stressful life events. *Psychological Bulletin, 136*(2), 257–301. https://doi.org/10.1037/a0018301

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163. https://doi.org/10.2307/271063

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173–184. https://doi.org/10.1177/01466216970212006

Richardson, C. G., Ratner, P. A., & Zumbo, B. D. (2007). A test of the age-based measurement invariance and temporal stability of Antonovsky's Sense of Coherence Scale. *Educational and Psychological Measurement, 67*(4), 679–696. https://doi.org/10.1177/0013164406292089

Rivera, F., García-Moya, I., Moreno, C., & Ramos, P. (2013). Developmental contexts and sense of coherence in adolescence: A systematic review. *Journal of Health Psychology, 18*(6), 800–812. https://doi.org/10.1177/1359105312455077

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561–582. https://doi.org/10.1080/10705510903203433

Satorra, A. (2000). Scaled and adjusted restricted tests in multisample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp. 233–247). Kluwer Academic.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210–222. https://doi.org/10.1016/j.hrmr.2008.03.003

Schumacher, J., Wilz, G., Gunzelmann, T., & Brähler, E. (2000). Die Sense of Coherence Scale von Antonovsky – Teststatistische Überprüfung in einer repräsentativen Bevölkerungsstichprobe und Konstruktion einer Kurzskala [The Antonovsky Sense of Coherence Scale. Test statistical evaluation in a representative population sample and construction of a brief scale]. *Psychotherapie, Psychosomatik, Medizinische Psychologie, 50*(12), 472–482. https://doi.org/10.1055/s-2000-9207

Schumacker, R. E., & Lomax, R. G. (2010). *A beginners guide to structural equation modeling*. Routledge.

Silverstein, M., & Heap, J. (2015). Sense of coherence changes with aging over the second half of life. *Advances in Life Course Research, 23*, 98–107. https://doi.org/10.1016/j.alcr.2014.12.002

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78–90. https://doi.org/10.1086/209528

Stein, J. A., Lee, J. W., & Jones, P. S. (2006). Assessing cross-cultural differences through use of multiple-group invariance analyses. *Journal of Personality Assessment, 87*(3), 249–258. https://doi.org/10.1207/s15327752jpa8703_05

Stern, B., Socan, G., Rener-Sitar, K., Kukec, A., & Zaletel-Kragelj, L. (2019). Validation of the Slovenian version of short Sense of Coherence Questionnaire (SOC-13) in multiple sclerosis patients. *Slovenian Journal of Public Health, 58*(1), 31–39. https://doi.org/10.2478/sjph-2019-0004

Tran, T. V. (2009). *Developing cross-cultural measurement invariance*. Oxford University Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–69. https://doi.org/10.1177/109442810031002

Zimprich, D., Allemand, M., & Hornung, R. (2006). Measurement invariance of the abridged Sense of Coherence Scale in adolescents. *European Journal of Psychological Assessment, 22*(4), 280–287. https://doi.org/10.1027/1015-5759.22.4.280

History
Received August 1, 2019
Revision received September 18, 2020
Accepted January 7, 2021
Published online May 26, 2021
EJPA Section / Category Positive Psychology & Assessment

Open Science
We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact *p* values, effect sizes, and 95% confidence or credible intervals.
Open Data: The information needed to reproduce all of the reported results are not openly accessible.
Open Materials: I confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology (Grevenstein, 2021).
Preregistration of Studies and Analysis Plans: This study was not preregistered.

Funding
Open access publication enabled by Heidelberg University.

ORCID
Dennis Grevenstein
https://orcid.org/0000-0002-1422-0309

Dennis Grevenstein
Psychological Institute
Heidelberg University
Hauptstr. 47-51
69117 Heidelberg
Germany
dennis.grevenstein@psychologie.uni-heidelberg.de

Erratum

Correction to Grevenstein & Bluemke, 2021

The article entitled "Measurement invariance of the SOC-13 Sense of Coherence Scale across gender and age groups" by Dennis Grevenstein & Matthias Bluemke (*European Journal of Psychological Assessment*, 1–11, https://doi.org/10.1027/1015-5759/a000641) has now been published as an open access article with "© The Author(s)" and under a CC BY-NC 4.0 license.

The following funding information has been added:

Funding
Open access publication enabled by Heidelberg University.

Reference

Grevenstein, D., & Bluemke, M. (2021). Measurement invariance of the SOC-13 Sense of Coherence Scale across gender and age groups. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000641

Published online January 27, 2022

Instructions to Authors

The main purpose of the *European Journal of Psychological Assessment* is to present important articles, which provide seminal information on both theoretical and applied developments in this field. Articles reporting the construction of new measures or an advancement of an existing measure are given priority. The journal is directed to practitioners as well as to academicians: The conviction of its editors is that the discipline of psychological assessment should, necessarily and firmly, be attached to the roots of psychological science, while going deeply into all the consequences of its applied, practice-oriented development.

Psychological assessment is experiencing a period of renewal and expansion, attracting more and more attention from both academic and applied psychology, as well as from political, corporate, and social organizations. The *EJPA* provides a meeting point for this movement, contributing to the scientific development of psychological assessment and to communication between professionals and researchers in Europe and worldwide.

European Journal of Psychological Assessment **publishes the following types of articles:** Original Articles, Brief Reports, Multistudy Reports, and Registered Reports.

Manuscript submission: All manuscripts should in the first instance be submitted electronically at http://www.editorialmanager.com/ejpa. Detailed instructions to authors are provided at http://www.hgf.io/ejpa

Copyright Agreement: By submitting an article, the author confirms and guarantees on behalf of him-/herself and any coauthors that the manuscript has not been submitted or published elsewhere, and that he or she holds all copyright in and titles to the submitted contribution, including any figures, photographs, line drawings, plans, maps, sketches, tables, and electronic supplementary material, and that the article and its contents do not infringe in any way on the rights of third parties. ESM will be published online as received from the author(s) without any conversion, testing, or reformatting. They will not be checked for typographical errors or functionality. The author indemnifies and holds harmless the publisher from any third-party claims.

The author agrees, upon acceptance of the article for publication, to transfer to the publisher the exclusive right to reproduce and distribute the article and its contents, both physically and in nonphysical, electronic, or other form, in the journal to which it has been submitted and in other independent publications, with no limitations on the number of copies or on the form or the extent of distribution. These rights are transferred for the duration of copyright as defined by international law. Furthermore, the author transfers to the publisher the following exclusive rights to the article and its contents:

1. The rights to produce advance copies, reprints, or offprints of the article, in full or in part, to undertake or allow translations into other languages, to distribute other forms or modified versions of the article, and to produce and distribute summaries or abstracts.
2. The rights to microfilm and microfiche editions or similar, to the use of the article and its contents in videotext, teletext, and similar systems, to recordings or reproduction using other media, digital or analog, including electronic, magnetic, and optical media, and in multimedia form, as well as for public broadcasting in radio, television, or other forms of broadcast.
3. The rights to store the article and its content in machine-readable or electronic form on all media (such as computer disks, compact disks, magnetic tape), to store the article and its contents in online databases belonging to the publisher or third parties for viewing or downloading by third parties, and to present or reproduce the article or its contents on visual display screens, monitors, and similar devices, either directly or via data transmission.
4. The rights to reproduce and distribute the article and its contents by all other means, including photomechanical and similar processes (such as photocopying or facsimile), and as part of so-called document delivery services.
5. The right to transfer any or all rights mentioned in this agreement, as well as rights retained by the relevant copyright clearing centers, including royalty rights to third parties.

Online Rights for Journal Articles: If you wish to post the article to your personal or institutional website or to archive it in an institutional or disciplinary repository, please use either a pre-print or a post-print of your manuscript in accordance with the publication release for your article and the document "Guidelines on sharing and use of articles in Hogrefe journals" on the journal's web page at http://www.hgf.io/ejpa

January 2021

APPLICATION FORM

EAPA membership includes a free subscription to the *European Journal of Psychological Assessment.* To apply for membership in the EAPA, please fill out this application form and return it together with your *curriculum vitae* to: David Gallardo-Pujol, PhD (EAPA Secretary General), Dept. of Clinical Psychology & Psychobiology, Campus Mundet, Pg. de la Vall d'Hebron, 171, 08035 Barcelona, Spain, E-mail secretary@eapa.science.

Family name . . . . . .

First name . . . . . .

Affiliation . . . . . .

Address . . . . . .

City . . . . . Postcode . . . . . .

Country . . . . . .

Phone . . . . . Fax . . . . . .

E-mail . . . . . .

ANNUAL FEES

◆ EURO 75.00 (US $ 98.00) – Ordinary EAPA members

◆ EURO 50.00 (US $ 65.00) – PhD students

◆ EURO 10.00 (US $ 13.00) – Undergraduate student members

IMPORTANT!
3-digit security code in signature field on reverse of card (VISA/Mastercard) or 4 digits on the front (AmEx)

FORM OF PAYMENT

◆ *Credit card*

☐ VISA ☐ Mastercard/Eurocard ☐ American Express

Number ☐☐☐☐ ☐☐☐☐ ☐☐☐☐ ☐☐☐☐

Expiration date ☐☐ / ☐☐ CVV2/CVC2/CID# ☐☐☐ ☐

Card holder's name . . . . . .

Signature . . . . . Date . . . . . .

◆ *Cheque or postal order*

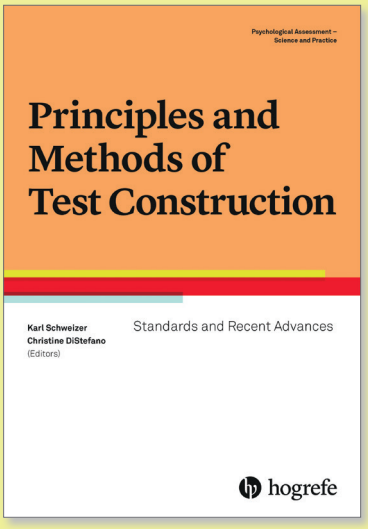
Send a cheque or postal order to the address given above

Signature . . . . . Date . . . . . .

Test development and construction: Current practices and advances

"This book is indispensable for all who want an up-to-date resource about constructing valid tests."

Prof. Dr. Johnny R. J. Fontaine, President of the European Association of Psychological Assessment, Faculty of Psychology and Educational Sciences, Ghent University, Belgium



Karl Schweizer / Christine DiStefano (Editors)

Principles and Methods of Test Construction

Standards and Recent Advances

Psychological Assessment – Science and Practice, vol. 3)
2016, vi + 336 pp.
US $69.00 / € 49.95
ISBN 978-0-88937-449-2
Also available as eBook

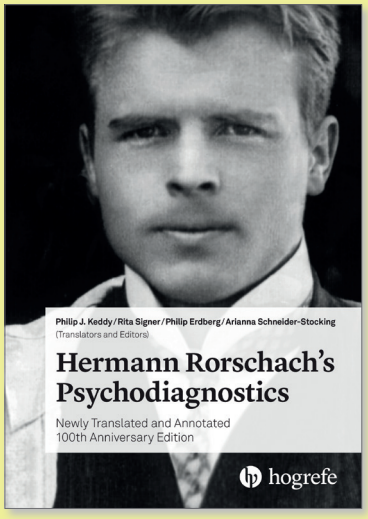
This volume in the series *Psychological Assessment – Science and Practice* describes the current state-of-the-art in test development and construction. The past 10-20 years have seen substantial advances in the methods used to develop and administer tests. In this volume many of the world's leading authorities collate these advances and provide information about current practices, thus equipping researchers and students to successfully construct instruments using the latest standards and techniques. The volume is organized into five related sections. The first explains the benefits of considering the underlying theory when designing tests, with a focus on factor analysis and item response theory in construction. The second section looks at item format and test presentation. The third discusses model testing and selection, while the fourth goes into statistical methods to identify group-specific biases. The final section discusses current topics of special relevance, such as multitrait multi-state analyses and development of screening instruments.

www.hogrefe.com

hogrefe

A celebration of Hermann Rorschach's seminal text with this completely new translation!

100 Years Rorschach® Test



Philip J. Keddy / Rita Signer / Philip Erdberg / Arianna Schneider-Stocking (Translators and Editors)

Hermann Rorschach's Psychodiagnostics

Newly Translated and Annotated 100th Anniversary Edition

2021, xviii + 294 pp.
US $69.00 / € 59.95
ISBN 978-0-88937-558-1
Also available as eBook

This new English translation and 100th anniversary annotated edition of *Psychodiagnostics,* the only book published by Hermann Rorschach, showcases Rorschach's empiricism and the wide-ranging flexibility of his thinking – and thus helps us to understand why his iconic inkblot test has survived for a century and is still being used around the world, with the support of a strong evidence base. The expert translation team have collaborated closely to create an accessible rendition of Hermann Rorschach's presentation of the inkblot test that resulted from his empirical research experiments. Also included is the case study lecture that Rorschach gave to the Swiss Psychoanalytic Society in 1922, just six weeks before his premature death. Both his book and the lecture are accompanied by annotations for the first time, looking backward to the sources of Rorschach's terminology and also forward to how the test is used today. Drawings and photographs from the Rorschach Archive as well as introductory chapters on the history of the translation and the creation of *Psychodiagnostics* bring the story of this important figure and his work to life. This volume is essential reading for both historians and contemporary users of the inkblot test and anyone interested in exploring personality testing.

www.hogrefe.com

hogrefe