SURVEYS · QUESTIONNAIRE DESIGN

The Development and Validation of Short-Form Survey Instruments

Short forms are reduced versions of longer questionnaires and scales. This article explains why they are built, how items are chosen, what is lost when a scale is shortened and how a short form should be validated before it is trusted.

ModQu research article · 2,800 words · 26 references (APA 7th edition)

Many of the best-known measures in the social and health sciences began as long instruments. Personality inventories once ran to hundreds of items, and general health questionnaires to dozens. Long instruments measure precisely, but they take time, and every minute spent on one measure is a minute that cannot be spent on another. Survey designers therefore often turn to **short forms**: reduced versions of established instruments that aim to measure the same thing with fewer questions.

Short forms are everywhere. The 12-item health survey known as the SF-12 was derived from the 36-item SF-36 (Ware et al., 1996). The two-item Patient Health Questionnaire (PHQ-2) screens for depression using the first two items of the nine-item PHQ-9 (Kroenke et al., 2003). Personality can be measured with ten items rather than forty-four (Rammstedt & John, 2007). Yet methodologists have warned for decades that short forms are often built and used carelessly, and that shortening a scale can damage it in ways that are not obvious (Levy, 1968; Smith et al., 2000). This article explains how to develop a short form well and how to check that it works.

Why shorten an instrument?

The main reason is respondent burden. Long questionnaires reduce participation and the quality of answers, especially in online surveys and in panels where the same people are asked repeatedly. A shorter measure allows a survey to include more topics, to reach respondents who would not complete a long questionnaire, or to be repeated more often.

Cost is a second reason. In interviewer-administered surveys, each additional minute adds to fieldwork costs. In clinical settings, staff time is limited, and a two-item screener can be used routinely where a longer instrument could not (Kroenke et al., 2003).

A third reason is the context of use. Large multi-topic surveys, such as household panels, often need to include personality, well-being or health as one module among many. Very short measures of personality were developed for exactly this situation (Gosling et al., 2003; Rammstedt & John, 2007).

Well-known short forms and how they were made

The history of short forms shows several different routes. The SF-36 was itself a shortened set of items drawn from the much longer instruments of the Medical Outcomes Study, chosen to cover eight health concepts in a questionnaire that could be completed in a few minutes (Ware & Sherbourne, 1992). Its developers later used regression methods to select the 12 items that best reproduced the two summary scores of the SF-36, physical and mental health, creating the SF-12 (Ware et al., 1996). The SF-12 was therefore designed to support the summary scores, not all eight subscales, which is a good example of a short form tied to a stated purpose.

Personality research took a different route. The Ten-Item Personality Inventory was written as a new very brief measure, with two items for each of the five broad traits, rather than selected from an existing scale (Gosling et al., 2003). The Mini-IPIP, by contrast, selected four items per trait from a longer public-domain inventory and was tested in several samples, including a test of whether it predicted the same outcomes as the longer version (Donnellan et al., 2006). The BFI-10 selected two items per trait from the 44-item Big Five Inventory and was evaluated in English and German samples (Rammstedt & John, 2007).

These examples illustrate an older debate about how scales should be constructed in the first place. Burisch (1984) compared rational, intuitive and statistical approaches to building personality inventories and found that short scales written from a clear definition of the construct often performed as well as long scales built by statistical selection. The lesson for short forms is that careful thought about content can matter as much as statistical optimisation.

What is lost when a scale is shortened

Reliability

Reliability, in the sense of internal consistency, depends partly on the number of items. Other things being equal, a scale with fewer items is less reliable. The relationship is described by the Spearman–Brown formula, which predicts how reliability changes when a test is lengthened or shortened with items of the same quality (Spearman, 1910; Brown, 1910). Figure 1 shows the predicted pattern for a 20-item scale with a reliability of 0.90.

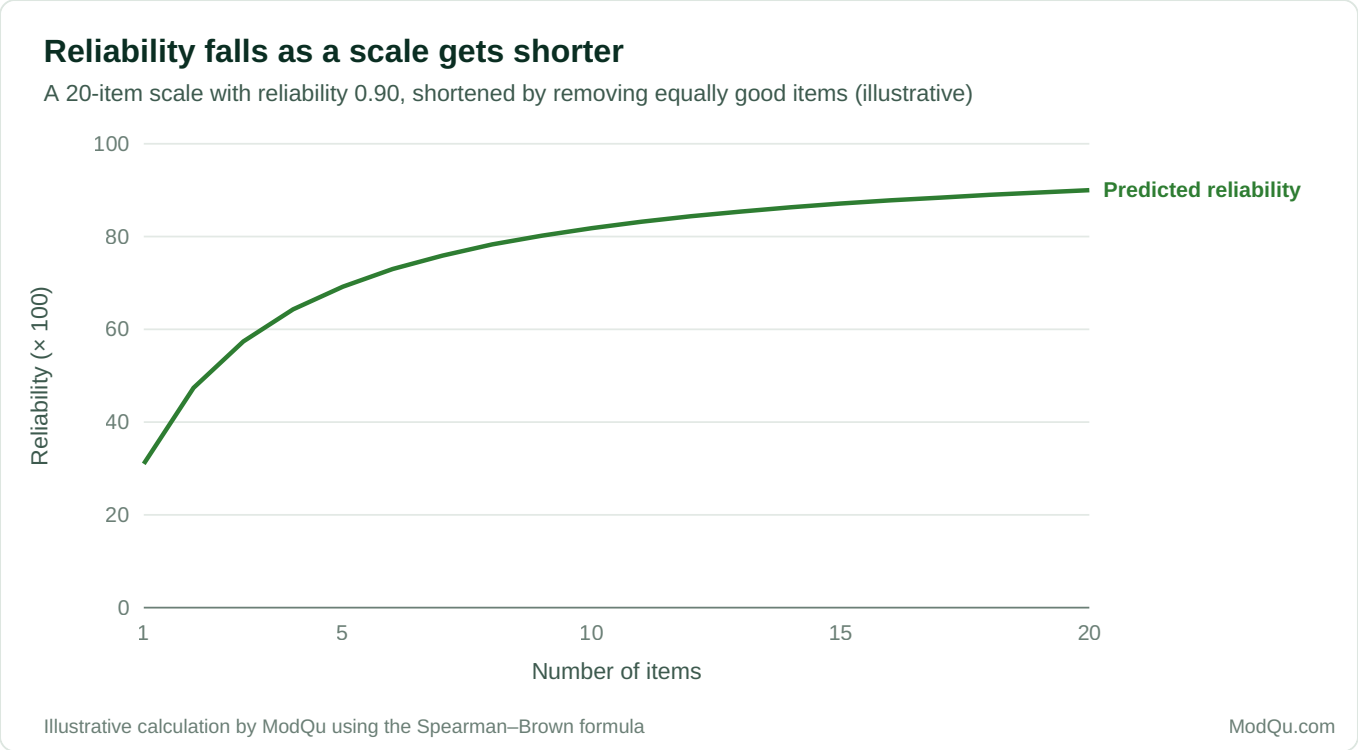


Figure 1. Predicted reliability of a scale as items are removed, for an illustrative 20-item scale with reliability 0.90. Illustrative calculation by ModQu using the Spearman–Brown formula.

The curve is flat at the top and steep at the bottom. Halving the scale from 20 to 10 items reduces reliability only modestly, from 0.90 to about 0.82. Going from 10 items to 2 reduces it much more, to below 0.50. In practice, developers can do better than the formula predicts by keeping the strongest items, but the general lesson holds: the first cuts are cheap and the last ones are expensive.

Precision for individuals

Lower reliability means larger measurement error for each respondent. For estimates of group averages this may matter little, because errors cancel out across large samples. For decisions about individuals, such as classifying a patient or selecting an employee, it can matter a great deal. A review of shortened tests concluded that many short forms are too unreliable for decisions about individuals, even when they are adequate for research on groups (Kruyen et al., 2013).

Content coverage

The less obvious loss is breadth. Most constructs have several facets. A measure of job satisfaction, for example, may cover pay, colleagues, supervision and the work itself. If items are chosen only because they correlate highly with each other, the short form may end up measuring a narrow part of the construct very consistently while ignoring the rest. This is sometimes called the attenuation paradox: increasing internal consistency by keeping similar items can reduce validity (Clark & Watson, 1995; Smith et al., 2000).

Validity

A short form can be reliable and still not measure the same thing as the full instrument, or predict outcomes as well. Comparisons of short and long measures of personality found that very short versions predicted outcomes less well and could lead to wrong conclusions about which traits mattered (Credé et al., 2012). The consequences of shortening therefore need to be tested, not assumed.

How items are selected

Statistical criteria

The most common approach is to keep the items that perform best in data from the full instrument. Typical criteria include the correlation of each item with the total score, the loading of each item on its factor in factor analysis, and the item's contribution to Cronbach's alpha (Cronbach, 1951; Stanton et al., 2002).

Item response theory offers more refined criteria. Each item's discrimination and difficulty are estimated, and the developer can choose items that together give precise measurement over the range of the trait that matters for the intended use (Edelen & Reeve, 2007). A short form for screening, for example, should measure precisely near the cut-off, while a short form for population research may need precision across the whole range. Short forms drawn from calibrated item banks, such as those of PROMIS, use this logic (Cella et al., 2010).

Automated search methods have also been used. Because the number of possible subsets of items is very large, algorithms can search for the subset that best meets several criteria at once, such as reliability, model fit and correlation with the full scale (Leite et al., 2008).

Content criteria

Statistics alone are not enough. Developers should first decide which facets the short form must cover and then select items within each facet. Experts in the subject, and sometimes members of the target population, can judge whether the chosen items still represent the construct. Marsh et al. (2005) proposed explicit criteria for short-form evaluation that combine statistical fit with the preservation of the original structure.

Balance and wording

Practical points also matter. If the original scale balanced positively and negatively worded items to control for acquiescence bias, the short form should keep that balance (Rammstedt & John, 2007). Items that depend on other items for their meaning, or that refer to a time frame set by an earlier question, may not work on their own.

Validating a short form

Smith et al. (2000) described a set of common errors in short-form development, which they called the "sins" of short-form development. Their main points, together with later reviews, suggest a clear process for validation (Kruyen et al., 2013). Figure 2 summarises it.

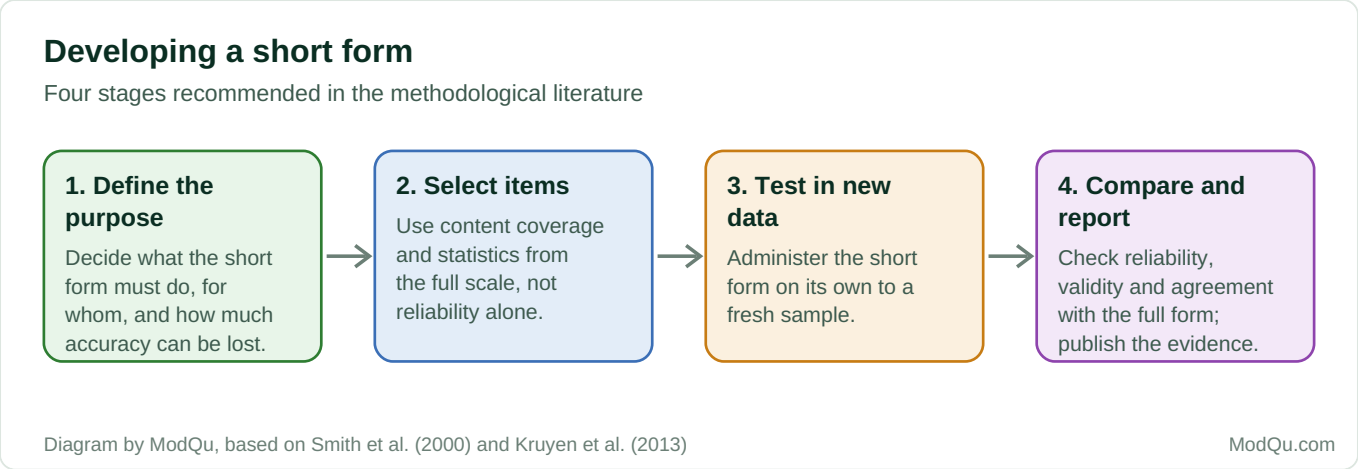


Figure 2. Stages in developing and validating a short form. Diagram by ModQu, based on Smith et al. (2000) and Kruyen et al. (2013).

Use new data

A short form chosen using one dataset will look better in that dataset than it really is, because the item selection takes advantage of chance features of the sample. This is a form of overfitting. The short form should therefore be tested in a separate sample (Smith et al., 2000).

Administer the short form on its own

Many short forms are evaluated by extracting the chosen items from responses to the full instrument. This is convenient but can mislead. When items are answered as part of a long scale, each answer is influenced by the context of the other items. When the short form is administered on its own, that context is missing, and the items may behave differently. Proper validation requires a study in which the short form is answered by itself (Smith et al., 2000).

Check reliability and structure

The short form's reliability should be estimated and reported, and its factor structure should be checked to see whether it matches that of the full instrument. Where the full instrument has several subscales, it is important to know whether each subscale still works or only the total score is trustworthy (Marsh et al., 2005).

Compare with the full form and with outcomes

The short form's scores should correlate highly with the full form's scores, but this correlation is inflated when the short form's items are part of the full form, because the same answers appear in both. Correlations should therefore be corrected or estimated with separate administrations. More importantly, the short form should be tested against outcomes and other measures that the full instrument predicts, to show that it supports the same conclusions (Messick, 1995; Credé et al., 2012).

State the purpose

Finally, a short form should be validated for its intended use. A screener needs evidence about how well it identifies cases, such as sensitivity and specificity at a cut-off. A research measure needs evidence about group comparisons and associations. Evidence for one purpose does not automatically justify another (Kruyen et al., 2013).

Single-item measures

The extreme case of a short form is a single question. Single items cannot have internal consistency, so their reliability must be judged in other ways, such as repeated measurement. They have a mixed record. For constructs that are concrete and clearly understood by respondents, such as overall job satisfaction or liking of an advertisement, single items can perform well (Wanous et al., 1997; Bergkvist & Rossiter, 2007). For abstract or multi-faceted constructs, they usually do not. A study of predictive validity concluded that single items should be used only under narrow conditions, such as when the construct is concrete, the sample is small and the cost of extra items is high (Diamantopoulos et al., 2012).

Many business surveys rely on single items, the Net Promoter Score being a well-known example. These can be useful for tracking, but they should not be treated as precise measures of complex attitudes.

Short forms and adaptive testing

Short forms and computerised adaptive testing are two answers to the same problem. A short form gives every respondent the same small set of items. An adaptive test chooses items for each respondent based on earlier answers, so that each person receives the items that are most informative for them. When an item bank calibrated with item response theory exists, both can be built from it. Static short forms are easier to administer, work on paper and in any survey software, and produce scores that are simple to compare. Adaptive tests can reach the same precision with fewer items, but they require specialised software and calibrated banks (Edelen & Reeve, 2007; Cella et al., 2010). For most general-purpose surveys, a well-validated static short form remains the practical choice.

Reporting results from short forms

Results based on short forms should be reported with the same care as the measures themselves. Reports should name the short form and cite its validation study, state the reliability observed in the current sample and avoid comparing scores directly with norms established for the full instrument unless a link between the two has been demonstrated. Where the short form produces a total score only, reports should not present subscale results. When an organisation changes from a full instrument to a short form in a tracking study, the change should be noted, and if possible both versions should be run in parallel for one wave, so that any shift in results can be separated from a real change in the population.

Common misunderstandings

Some researchers reject short scales entirely because they are less reliable than the full versions. Ziegler et al. (2014) argued that this is too simple. Reliability is only one property, and a short scale with somewhat lower internal consistency can still be valid for its purpose, especially when it covers the construct's breadth. Very high internal consistency in a short scale may even be a warning sign that the items are near-duplicates.

The opposite mistake is to assume that a published short form is valid because the full form is. Each short form is a new instrument and needs its own evidence (Smith et al., 2000).

An example

A company runs an annual employee survey that includes a 24-item engagement scale with four facets: energy, dedication, involvement and pride. The full scale takes about six minutes, and the company wants a version for quarterly pulse surveys that takes under two minutes. The research team takes the following steps (illustrative example):

1. It agrees that the short form will be used only to track average engagement across departments, not to assess individuals.
2. Using last year's data, it selects two items from each facet, choosing those with the highest loadings while checking that the wording covers the facet's meaning.
3. It splits the data into two halves, selects items in the first half and checks reliability and correlation with the full scale in the second.
4. In the next quarter, it administers the eight-item short form on its own to a random half of employees and the full scale to the other half, comparing the distributions, reliabilities and relationships with turnover intentions.
5. It documents the results, including the expected loss of precision, so that future users know what the short form can and cannot support.

In business surveys

Short forms are especially attractive in commercial research, where questionnaires compete for limited respondent time and clients want many topics covered. The same principles apply. Borrowing a few items from an academic scale, changing their wording and calling the result a measure of the construct creates an untested instrument. Where an established short form exists, it is usually better to use it unchanged. Where a new short form is needed, the investment in a proper validation study is small compared with the cost of decisions based on a measure that does not work.

Checklist for developers

- Define the construct and its facets before selecting items.
- State the purpose of the short form and the precision required.
- Use both statistical criteria and expert judgement to choose items.
- Keep the balance of item wording where the original had it.
- Test the short form in a new sample, administered on its own.

- Report reliability, structure, correlation with the full form and relationships with outcomes.
- Publish the evidence and the intended use, so others can judge whether it suits their purpose (Boateng et al., 2018).

Checklist for users

- Check whether the short form has been validated for a purpose like yours.
- Use the published items and response scale without changes.
- Remember that individual scores are less precise than group averages.
- Avoid interpreting subscales from a short form unless they were shown to work.

Conclusion

Short-form instruments allow surveys to measure more with less, and some of the most widely used measures in research and practice are short forms. Shortening, however, is not a neutral act. It reduces reliability, can narrow content and can change what is measured. Good short forms are built deliberately, with a clear purpose, a balance between statistics and content, and validation in new data where the short form stands on its own. When these steps are followed, short forms can provide sound knowledge at a fraction of the burden. When they are skipped, the savings in respondent time can be outweighed by the cost of measuring the wrong thing (Smith et al., 2000; Kruyen et al., 2013).

References

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*(2), 175–184. https://doi.org/10.1509/jmkr.44.2.175

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39*(3), 214–227. https://doi.org/10.1037/0003-066X.39.3.214

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., … Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102*(4), 874–888. https://doi.org/10.1037/a0027403

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

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

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(Suppl. 1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2003). The Patient Health Questionnaire-2: Validity of a two-item depression screener. *Medical Care, 41*(11), 1284–1292. https://doi.org/10.1097/01.MLR.0000093487.78664.3C

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing, 13*(3), 223–248. https://doi.org/10.1080/15305058.2012.703734

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research, 43*(3), 411–431. https://doi.org/10.1080/00273170802285743

Levy, P. (1968). Short-form tests: A methodological review. *Psychological Bulletin, 69*(6), 410–416. https://doi.org/10.1037/h0025736

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analytic methods. *Psychological Assessment, 17*(1), 81–102. https://doi.org/10.1037/1040-3590.17.1.81

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102–111. https://doi.org/10.1037/1040-3590.12.1.102

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*(1), 167–194. https://doi.org/10.1111/j.1744-6570.2002.tb00108.x

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82*(2), 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483. https://doi.org/10.1097/00005650-199206000-00002

Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care*, *34*(3), 220–233. https://doi.org/10.1097/00005650-199603000-00003

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales: Five misunderstandings and ways to overcome them. *Journal of Individual Differences, 35*(4), 185–189. https://doi.org/10.1027/1614-0001/a000148