Developing, evaluating and using subjective scales of personality, preferences, and well-being: A guide to psychometrics for psychologists and economists

Alex M. Wood & Christopher J. Boyce, Behavioural Science Centre, Stirling Management School, University of Stirling

Full reference: Wood, A. M., & Boyce, C. (2017). Developing, evaluating and using subjective scales of personality, preferences, and well-being: A guide to psychometrics for psychologists and economists. In Ranyard, R. (Ed). *Economic psychology: the science of economic mental life and behaviour* (pp. 88-105). Chichester, UK: Wiley.

Outline

6.1 Introduction

6.2 The Importance of Psychometrics for Economic Psychology Research

6.3 Steps in Developing a Scale

6.4 Other Steps and Conclusion

6.5 Summary

6.1.Introduction

Subjective measures of well-being, attitudes, psychological states, and personality are increasingly used in economic research (Borghans, Duckworth, Heckman, & Weel, 2008; Dolan, Peasgood, & White, 2008). They have not only played an important role in overturning traditional economic models (Kahneman, Wakker, & Sarin, 1997) but have also helped our understanding of how individuals react to changing socio-demographic circumstances (Boyce & Wood, 2011; A. E. Clark, Diener, Georgellis, & Lucas, 2008). This burgeoning economic research has introduced many advanced statistical econometric techniques to psychological research and this has helped establish important causal relationships and yield potentially important policy conclusions (O'Donnell, Deaton, Durand, Halpern, & Layard, 2014). However, the credibility of these results depend on how accurately and with what precision variables are measured (reliability) and whether the variable being measured is that which is claimed to being measured (validity). Analogically, there would be little value to an astronomer in having a poorly calibrated telescope pointed at the right galaxy (low reliability, high validity) or a well calibrated telescope pointed at the incorrect galaxy (high reliability, low validity).

Psychometrics, an area of psychology that has developed to specifically ensure psychological constructs are measured with optimum reliability and validity, is virtually unheard of in economics. Psychometrics is also a specialisation within psychology, and there

is a need for greater dissemination amongst psychology students who use but don't design scales. In this chapter we therefore offer a guide to the necessary steps in developing subjective measures[1] and in doing so we hope to introduce researchers to psychometrics, enabling them to evaluate measures used by others and use appropriate measures in their own research.

6.2 The Importance of Psychometrics for Economic Psychology Research

Psychometrics as a field essentially reduces to a set of techniques designed to construct a subjectively reported instrument with optimum reliability and validity (and to test the performance of existing instruments on these two dimensions). Fifty years ago within psychology, it would be common to use instruments in studies which had not been developed or tested according to the then poorly understood psychometric principles. However over the last few decades it has become near unthinkable to publish in a reputable journal a paper reporting instruments which have *not* undergone such testing. As the study of subjective constructs within economics is far newer, and there hasn't been dissemination of psychometric practice within economics, understandably practice is as it was historically in psychology. A consequence of this is that there are often inappropriate or un-validated scales contained within large panel datasets which likely do not measure what they purport to measure, or at the very least do not do so as well as they might have done had they been developed according to psychometric principles. Further, if the importance of scale development is not fully understood then researchers may make the mistake of confusing one measurement with another just because it "appears" to be the same.

[1] Note on terminology. Following conventions within psychology, subjective reports refer to a person answering questions about themselves (self-report) or others (peer-report) on some psychological characteristics (e.g., well-being, attitudes, traits, states etc.). One or (usually) more questions purporting to measure the same construct are referred to as a scale, measure, or instrument; if sub-scores are formed within a scale these are referred to as sub-scales.

One informative example is that of subjective measures of well-being. For example, life satisfaction and positive affect are often interchangeably referred to as “happiness”, and this has caused great confusion in the literature. However, life satisfaction, representing an individual’s evaluation of their life overall, is very different from how an individual feels in a given moment (affect). It has been shown, for example, that life satisfaction and positive affect have different relationships with both income (Kahneman & Deaton, 2010) and unemployment (Knabe, Rätzel, Schöb, & Weimann, 2010). In developing a scale it is therefore essential to have the right meaning of the construct represented in the items or any conclusions based on its use risk being invalid. Given the suggestion to evaluate the progress of society based on national indicators of subjective well-being alongside more traditional economic indicators such as GDP growth (O’Donnell et al., 2014; Stiglitz, Sen, & Fitoussi, 2009) it is important to ensure we understand exactly what is being measured. Whilst it would be unthinkable for the national policies of OECD countries to be based on economic data that could not be verified, the subjective measures being collected do not meet the most rudimentary psychometric standards.

The use of subjective measures within economics, whilst fast growing, is nevertheless still a fringe activity and we routinely encounter scepticism as to the appropriateness of using such measures (see e.g., Johns & Ormerod, 2012). To the extent that these refer to measures that have not been psychometrically validated then we would share this scepticism. But to the extent that these criticisms refer to the practical possibility of measuring subjective constructs, they simply represent a lack of knowledge of psychometrics and the huge amount of validation that has overwhelmingly supported the reliability and validity of the most validated scales. Thus knowledge of psychometrics allows an evidence based response to criticism of the use of subjective measures within economics, and the widespread knowledge

of psychometrics in the field would allow such criticism to be more appropriately targeted at specific uses rather than generalisations about the endeavour.

We stress that we do not mean to be overly critical of existing work within economics that use such scales (which includes our own), as at the moment there is a trade-off between collecting one's own data (as commonly used by psychologists) with optimally psychometrically defined variables, and using existing very large (N > 5,000) representative longitudinal datasets (as commonly used by economists). The trade-off is often currently in favour of using these large datasets with non-optimal variables. However, even within these constraints, psychometrics can be used to test the performance of the measures in these datasets (either with the information contained within, or as small supplementary validation studies) and also help individual researchers more accurately assess the limitations of their research. In the longer term, we hope that wider knowledge of psychometrics within economics will encourage better scales to be included in new and revised surveys.

This chapter is the first to our knowledge to introduce psychometrics to an economics audience. Many excellent textbooks on psychometrics exist (e.g., Coaley, 2010) as do highly cited guides introducing psychometrics to sub-fields within psychology (e.g., Worthington & Whittaker, 2006), and excellent overviews of specific psychometric techniques (e.g., L. A. Clark & Watson, 1995; Hunsley & Meyer, 2003; Smith & McCarthy, 1995). We'd recommend these references are read alongside this chapter. However, the textbooks require substantial time investment for the typical user and this chapter therefore offers a clear overview of psychometrics that will not only help economists interpret research using subjective measures but will be useful for advanced psychology students interested in economic psychology.

We organise the remainder of this chapter around designing a new scale. A common question from colleagues within economics working with subjective measures is; "is this a good scale?" and the best way to evaluate a scale is to ask how many of the standard steps of psychometric development have been applied to the scale and how convincing the demonstration of each step is.

6.3 Steps in Developing a Scale

Table 1 gives examples of commonly used and well-validated psychometric measures from psychology that may be of interest to people at the interface of economics and psychology. As should be apparent they represent very different psychological constructs. The steps used in the development and evaluation of any such scales are however the same, as they would be for other self-report scales such as ones for use in marketing, business, or consumer psychology research. The basic underlying steps are outlined below and represent a toolkit that can be applied to any new problem at hand. Much of the information in this section is quite complex, and readers would not be expected to fully understand every last process described, but it is hoped that the basic message will be conveyed and the information here will act as a useful overarching guide.

[Table 1 about here]

Step 1: Identify the need for the scale

This step may appear obvious, but it is actually the one that is most omitted from most subjective scale development exercises. More than fifty years of work within psychology has resulted in subjective scales for pretty much whatever one would want to measure. Indeed, in the 1980s and 1990s a relatively easy way to get a publication was to develop a new scale; hence this was taken advantage of, and scale development was often suggested to PhD

students as a way to contribute to knowledge given the resources that they typically have. The benefit of this is that there are now a huge number of useful subjective scales available. However, the cost has been that multiple scales have emerged that measure similar but not quite identical constructs, often using slightly different terms for the construct and framing in terms of different prior literature. Such a multiplicity of scales has the advantage of being able to operationalise latent constructs with multiple scales, but has the disadvantage of making it more difficult for literature reviews to make substantive conclusions – are, for example, differences between studies due to substantive factors such as the populations studied, or are they simply due to the use of different scales? As a result of these concerns, publishing scale development papers has gone from on the easiest "wins" to amongst the most difficult category of papers to get published. Nevertheless, this is also the category of papers that generally gets the most citations – if a scale is genuinely needed and accepted as reliable and valid then it is likely to be highly used. It is notable that of the 100 most cited research papers of all time, the only entries from psychology are explicit scale development papers (Van Noorden, Maher, & Nuzzo, 2014). Occasionally there is a genuinely unstudied fundamental construct. Sometimes there is a need for a shorter version of a lengthy scale. In applied research, new developments in services sometimes need new scales designed to assess progress. For example, advances in genetics have created new genetic testing services, which provide such information as the likelihood of any children having the same disorder as the parent. To support individuals going through this process, new genetic counselling services emerged, and in order to inform health care management, a scale was developed to assess whether patients' needs were being met (McAllister, Wood, Dunn, Shiloh, & Todd, 2011). Even more commonly, there is a need to validate older scales that perhaps were developed before optimal practice became the norm; or, as is sadly the case for many scales in management consultancy, developed with commercial rather than scientific interests in

mind. Thus, although psychometrics is still a thriving discipline it is essential that a scale development exercise begins with a full systematic literature review to ensure that the scale doesn't already exist and to locate theoretically related scales (the latter will also be needed in later steps). A proper systematic review and strong rationale in the introduction will help ensure the contribution of the scale is apparent, and adherence to the remaining steps will ensure the scale is suitably reliable and valid. See Hinshaw (2009) for a starting point of how to conduct such a review.

Step 2: Define the Construct and Develop Items

To begin with, a large pool of items that covers the full terrain of the construct is developed. There will commonly be around 100 items, both to ensure that as many factors as are genuinely in the data have enough items out of which to emerge, and to allow deletion of worse performing items in the next step. A key challenge at this stage is how to map out the full terrain of the construct, and there are several ways in which this may be achieved, of varying levels of acceptability.

2.1. *Qualitative work.* This is an example of where mixed methods approach – almost unheard of in economics – is the most appropriate. Often it makes sense for the end user community to define the construct. Key decisions for the researcher would be the choice of community and the form of qualitative analysis. In the genetic counselling example, we asked people with the condition to describe what they wanted out of their treatment. Other approaches could have been used such as drawing the construct definition from the service, researchers, or an existing theoretical model, but this would go against the intent to have a "patient reported outcome measure (PROM)" and it was their experiences that were meant to be the point of the scale. Here it is clear that the researcher's conception is inappropriate as they are not part of that community, and the choice of participants and qualitative method

(interviews) was relatively mandated. However, even in cases where the researcher is part of the community there is benefit in using qualitative research to define the construct – not least as other approaches (such as one's intuition or previous work in the field) may not be fully inclusive of the construct. In a second example, a scale was designed to measure "pre-sleep cognitions", the thought that people have whilst falling asleep (it was subsequently shown that positive thoughts improved sleep whereas negative thoughts impaired it, Wood, Joseph, Lloyd, & Atkins, 2009). In this Step, the developers gave participants a voice activated Dictaphone and asked them to speak aloud whatever they were thinking. Excluding the most common theme (that they couldn't sleep due to speaking aloud into the Dictaphone), a quite exhaustive database of thoughts was collected that formed the basis of the item pool.

2.2. *Pre-Existing Complete Lists*. A reason to omit the qualitative stage is if there happen to be a full list of items already developed. This is rare, but was seen in the development of the dominant "Five Factor Model" of personality, where personality at the highest level of abstraction is represented by agreeableness, conscientiousness, extraversion, neuroticism and openness (John & Srivastava, 1999). These (and their attendant scale items) were developed through having participants rate every single adjective in the English language dictionary that described a person (excluding those that referred to skills or non-psychological differences). Hence, assuming that all words considered to describe a culturally important characteristic of a person become represented in language, this may be seen as a complete list.

2.3. *Existing Theoretical Model*. This is where practice can get non-optimal. Quite commonly, researchers design items around an existing conception (that, e.g., specifies that a trait or form of well-being has three factors). This can be acceptable, as long as it is made very clear in the paper and all subsequent research that the scale isn't designed to measure the construct, but rather a particular conception of it. This can be acceptable if the aim was simply to test predictions of the particular conception (for example, we developed a measure

of authenticity, to assess a conception of it arising from a particular counselling approach, for testing the theories associated with that approach, Wood, Linley, Maltby, Baliousis, & Joseph, 2008). However, if the approach is correct in predicting a certain number of factors that underlie the data, then there is no real reason not to develop items that cover the whole domain; if the theory is correct, these should naturally emerge anyway. This approach to item development can be problematic as once the scale is published, quite likely people will use it as a measure of the whole construct despite cautions not to do this in the development paper.

2.4. *Using Items from Existing Scales*. This is occasionally appropriate, as for example when shortening an existing scale or merging two existing scales (a good example of which is Griffiths et al., 2015). The key danger is the validity of the resultant measure it is totally reliant on the (possibly poorly derived) items in the existing scales.

2.5. *Researchers Using their Own Conception*. This is totally unacceptable, as this is based solely on the opinion of a small number of researchers (statistically from very certain backgrounds). Such subjectivity shouldn't have a part in scientific scale development.

2.6 *Finishing the Step*. At the outcome of the development phase the researchers would have developed a clear conception and a large pool of items that fully covers it. Quite how the items themselves have been written partially depends on the decisions above. Often it would have been appropriate for the item generation to be an integrative process between researchers and qualitative participants.

Step 3: Identifying the scale's structure and selecting items based on factor analysis.

After the full item pool has been developed, the researchers should give out the item pool to a large sample of participants from the final population who will be using the scale. Once the data are collected, it will be subjected to exploratory factor analysis (EFA). EFA

ideally has 450 participants, at which time the factor structures are so stable there is little point in having more. At a minimum, 150 participants are needed, although this is almost certainly too small and only justifiable if it is an exceptionally difficult to access population. Several excellent guides should be consulted (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Floyd & Widaman, 1995). This step involves seeing how many factors naturally underlie the data. EFA as a technique essentially looks at how correlations between items naturally group together, asks how many groups there are, and which items are most representative of that group. Determining the correct number of factors is critical and is not a subjective decision; rather, techniques such as parallel analysis test whether each potential factor explains more variance than would be expected had it arisen by chance, using normal criterions of significance. This (or the similar MAP technique) should always be used rather than older more subjective criteria (O'Connor, 2000; Velicer, Eaton, & Fava, 2000). Once it has been determined how many factors underlie the data, then the researcher is in a position to choose whether or not to have sub-scales. Again the decision in this step is not subjective; in this case the researcher would be obliged to form sub-scales to represent each substantive factor found in the data. The EFA should also be used to reduce the number of items from the large initial pool to the number required for the final scale. After deciding on how many items are needed the final items are chosen based on the "highest loading" items on the EFA for each factor, corresponding to each sub-scale, that is, the items which are most representative of each. Again this decision is not subjective; it is not up to the researcher to decide which the "best" items are, but rather the best items are those that are statistically shown to be the most representative of the construct. Part of the rationale for having so many items originally is to allow the ones that looked sensible but didn't perform well to be eliminated. Typically, four items per sub-scale is considered an optimum number. Shorter scales are more likely to be used in the field as they reduce participant burden, although there

will undoubtedly be a trade-off with establishing internal consistency (see below), which is normally higher with larger scales. If the improvement in internal consistency from moving from a four-item scale to a five-item scale is trivial, this will provide justification to use the lower number. The outcome of this step is that it produces the final scale, consisting of a low number of items either as a single scale, or as series of subscales. For example, if a measure of personality were being developed then giving out enough items to initially cover all differences between people, and conducting an EFA (which would have told you there were five factors), and then selecting the four most representative items in each factor would give a final 20 item scale with five sub-scales. The remaining steps here involve showing that this is reliable and valid, but this current step is one of the most important to get right, as in future steps only the reduced number of items will need to be given out. Once the factor structure and items are determined there is no going back.

Step 4: Confirming the Factor Structure

The next step involves giving out the scale to several new groups of participants which comprise the most likely user groups of the final scale (the same considerations on numbers apply as in the previous step). On these groups a confirmatory factor analysis (CFA) should be performed. CFA is a technique which differs from EFA in that it doesn't explore what factor structure is naturally in the data, but rather tests a prior conception (which here arises from your EFA). Ideally a "multi-group CFA", a technique which shows that the factor structure is stable across your groups, will be carried out and once this is complete other groups (such as gender and other demographics) can be assessed by combining the samples and splitting by gender. CFA will also ideally be used to test between the *a priori* factor solution and other number of factors that theory may have suggested. Once the factor structure is confirmed, the remaining steps test reliability and validity. Commonly these will

use the same samples as used for the CFA (as other measures would be typically collected alongside).

Step 5: Testing internal consistency (reliability)

This step involves showing, through Cronbach's alpha, that the items within each subscale correlate as expected. This alpha is the adjusted average inter-correlation between the items, and ranges from zero to one with broadly the same interpretation as a correlation co-efficient. In general values above .80 are excellent, as they indicate equivalence, but below .70 are generally deemed unacceptable as they indicate that the items are measuring too disparate concepts. We offer some caution with interpreting Cronbach's alphas since some scales with a low number of items (3 or less), which may be used to limit participant burden, can have low content overlap and therefore typically produce low Cronbach's alphas (Sijtsma, 2009).

Step 6: Testing Temporal Stability (reliability)

In this step the stability of the scale over time is assessed. The scale is given to participants at two time points and it needs to be shown (with the interclass coefficient) that neither the rank order nor the mean level of responses changes substantially over time (values above .80 are again excellent as they show equivalence). The time period to choose can be tricky. They should be far enough apart to ensure that participants do not simply remember and report what they first stated, but not so far away that meaningful change may have occurred. For many variables, two or four weeks would be appropriate; often one might, with different samples, assess both time periods (as consistent results would remove both concerns above). Ideally, one might show that the scale is sensitive to change when it would be expected to (e.g., that mean levels on a new depression reduce during psychotherapy relative to a control group who remain on the waiting list).

Step 7: Showing Face Validity

This step simply shows that people believe that the items make sense. This should be already justified if Step 1 has been carried out correctly, but this step can be formalised by taking the final items to focus groups of end users (perhaps the same as used in Step 1) or experts. This will avoid the complaint that the items don't make sense to experts or end-users later on.

Step 8: Show Criterion Validity

Here, it should be shown that the items correlate with other constructs that they are expected to. If there are any existing measures of the same construct then these should be used. However, if not, previous theory should be used to determine what the appropriate constructs are. For example, when a gratitude questionnaire was developed, the authors expected it to theoretically relate to well-being and social relationships, so this formed their criterion validity. It has to be argued convincingly in the introduction which constructs were chosen and why. It is also important at this stage to argue what is the magnitude of the correlations you expected to see and show that they are within this range in the results ($r > .30$ may be sufficient for very different constructs, $r > .70$ may be appropriate for identical constructs).

Step 9: Show discriminate validity

The converse of the last step; this involves showing that the scale does not correlate with what it is not meant to (this is a near zero correlation, rather than a negative correlation which would be evidence of convergent validity). For example, the social desirability scales (see Table 1) should not correlate with the new scale. Finding other measures with which the scale should not correlate is tricky and again depends on theory, which may in some cases be

present (e.g., in the development of the PANAS, see Table 1, there was the theoretical expectation that high activation positive and negative affect would not strongly correlate based on prior neurological findings, and their test of discriminate validity showed that this was the case). Sometimes an experiment can be conducted to show that your scale doesn't differ across groups (e.g., react to a mood induction, if it is not meant to capture moods); Bayesian statistical approaches can get around problems with trying to "prove the null hypothesis" in this fashion. Again, the argument needs to be made for whatever has been chosen, and since no correlation is ever zero, how small it would have to be to have discriminant validity ($r < .10$ would be a reasonable default). Of course, the convincingness of the argument is key; one cannot simply call all small correlations "divergent validity" and all large ones "convergent validity" as this would just capitalise on chance characteristics in the dataset.

Step 10: Test predictive validity

This step is an extension of Step 8, except here you are showing that the scale can predict some future outcome. This may be changes in an outcome over time within a longitudinal design, or it could be an objective behaviour. Theoretically, Step 9 could also be extended in this way, to show that the measure does not lead to what is not meant to (although this is not seen in the literature). The next Step may also be demonstrated longitudinally.

Step 11: Test incremental validity

In the first step it is argued that the measure was needed. This is directly tested in what is perhaps the most critical step. Here it is necessary to show that the scale can predict certain expected outcomes beyond existing scales. For example, if it is claimed that the new improved scale of child behaviour has been developed then it is necessary to show (e.g.,

through multiple regression) that it can predict some outcome (perhaps objective behaviour) above and beyond the existing childhood behaviour scales. If it does, then this provides direct evidence that it is needed; if it does not, the evidence suggests against the need for the scale. You will have to pick your outcome with care and justify it (as any scale wouldn't be expected to incrementally predict all outcomes) and equally the choice of what to predict needs to justify that a "tough" test was set by controlling for the theoretically or literally competing scales and constructs. A scale should convince no one if shown to predict outcomes beyond known, but poor predictors, of that outcome (e.g., gender or income). If, however, there is success in this step, then there is empirical demonstration that the scale is a novel contribution.

6. 4 Other Steps and Conclusion

The steps above are the practical minimum of what one needs to show for psychometric development. Depending on the scale, other steps may be appropriate. For example, where there are predictions that mean levels of the scale will differ between groups (as in between clinical and non-clinical participants on a well-being measure), this should be shown directly. An "ROC Curve" analysis may be performed to find the optimum score on the scale to distinguish between groups. Convergence between a person rating themselves on the measure and others rating them may be appropriate. As can be seen, scale development is a complex and very detailed process. One can readily understand psychologists' dismissive attitude of scales that have not been through this. However, if a scale development has met all of these steps, one can have considerable confidence that it is accurately measuring what it claims. Greater dissemination of this knowledge within economics combined with ensuring that only scales are used which have met these steps will help resolve the controversy of using subjective measures within economics. End users of the research will have a higher

confidence that the research based on the use of these scales is as valid as those using traditional economic indicators.

6.5 Summary

Subjective measures are increasingly used in economic research yet credibility of the results from research relying on subjective measures depends on how accurately and with what precision variables are measured (reliability) and whether the variable being measured is that which is claimed to being measured (validity). Here we overview the essential components of developing psychometrically valid scales thus enabling researchers to evaluate the research that uses subjective scales and use appropriate measures in their own research.

Review questions

Imagine that you were designing a scale to measure employee well-being.

- How would you develop the item pool?
- Who would you use as participants for the EFA and CFA (and, if different, to answer the questions below)?
- How would you test for temporal stability (including both stability and sensitivity to change)?
- How would you demonstrate face validity?
- How would you test criterion validity?
- How would you test predictive validity?
- How would you test discriminant validity?
- How would you test incremental validity?

Further reading

Coaley, K. (2010). *An Introduction to Psychological Assessment and Psychometrics*. Los Angeles: SAGE Publications Ltd.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*, 806–838.

References

Borghans, L., Duckworth, A. L., Heckman, J. J., & Weel, B. ter. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, *43*(4), 972–1059. http://doi.org/10.3368/jhr.43.4.972

Boyce, C. J., & Wood, A. M. (2011). Personality prior to disability determines adaptation agreeable individuals recover lost life satisfaction faster and more completely. *Psychological Science*, *22*(11), 1397–1402. http://doi.org/10.1177/0956797611421790

Buysse, D. J., Reynolds III, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213. http://doi.org/10.1016/0165-1781(89)90047-4

Carver, C. S. (1997). You want to measure coping but your protocol's too long: Consider the brief cope. *International Journal of Behavioral Medicine*, *4*(1), 92–100. http://doi.org/10.1207/s15327558ijbm0401_6

Clark, A. E., Diener, E., Georgellis, Y., & Lucas, R. E. (2008). Lags and leads in life satisfaction: A test of the baseline hypothesis. *The Economic Journal*, *118*(529), F222–F243. http://doi.org/10.1111/j.1468-0297.2008.02150.x

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

Coaley, K. (2010). *An Introduction to Psychological Assessment and Psychometrics*. Los Angeles: SAGE Publications Ltd.

Cohen, S. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), *The social psychology of health* (pp. 31–67). Thousand Oaks, CA, US: Sage Publications, Inc.

Cohen, S., Mermelstein, R., Kamarck, T., & Hoberman, H. M. (1985). Measuring the functional components of social support. In I. G. Sarason & B. R. Sarason (Eds.), *Social Support: Theory, Research and Applications* (pp. 73–94). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-94-009-5115-0_5

Costa, P. T. J., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE Handbook of Personality Theory and Assessment: Personality Measurement and Testing*. SAGE.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75. http://doi.org/10.1207/s15327752jpa4901_13

Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, *29*(1), 94–122. http://doi.org/10.1016/j.joep.2007.09.001

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. http://doi.org/10.1037/1040-3590.18.2.192

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. http://doi.org/10.1037/1082-989X.4.3.272

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299. http://doi.org/10.1037/1040-3590.7.3.286

Goldberg, D. P., & Williams, P. (1991). *A user's guide to the General Health Questionnaire*. University of London.

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

Griffiths, A. W., Wood, A. M., Maltby, J., Taylor, P. J., Panagioti, M., & Tai, S. (2015). The Development of the Short Defeat and Entrapment Scale (SDES). *Psychological Assessment*, No Pagination Specified. http://doi.org/10.1037/pas0000110

Hinshaw, S. P. (2009). For over a century, Psychological Bulletin has been the leading source of systematic review articles for the entire discipline of psychology. Editorial. *Psychological Bulletin*, *135*(4), 511–515. http://doi.org/10.1037/a0014869

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*(4), 446–455. http://doi.org/10.1037/1040-3590.15.4.446

John, O. P. (1990). The 'Big Five' factor taxonomy: Dimensions of personality in the natural language and questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York: Guilford Press.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (2nd ed., pp. 102–138).

Johns, H., & Ormerod, P. (2012). *Happiness, Economics and Public Policy*. Institute of Economic Affairs.

Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences*, *107*(38), 16489–16493. http://doi.org/10.1073/pnas.1011492107

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, *112*(2), 375–406. http://doi.org/10.1162/003355397555235

Knabe, A., Rätzel, S., Schöb, R., & Weimann, J. (2010). Dissatisfied with life but having a good day: Time-use and well-being of the unemployed*. *The Economic Journal*, *120*(547), 867–889. http://doi.org/10.1111/j.1468-0297.2009.02347.x

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, *46*(2), 137–155. http://doi.org/10.1023/A:1006824100041

McAllister, M., Wood, A. M., Dunn, G., Shiloh, S., & Todd, C. (2011). The Genetic Counseling Outcome Scale: a new patient-reported outcome measure for clinical genetics services. *Clinical Genetics*, *79*(5), 413–424. http://doi.org/10.1111/j.1399-0004.2011.01636.x

McCullough, M. E., Emmons, R. A., & Tsang, J.-A. (2002). The grateful disposition: A conceptual and empirical topography. *Journal of Personality and Social Psychology*, *82*(1), 112–127. http://doi.org/10.1037/0022-3514.82.1.112

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 396–402. http://doi.org/10.3758/BF03200807

O'Donnell, G., Deaton, A., Durand, M., Halpern, D., & Layard, R. (2014). Wellbeing and policy. Legatum Institute.

Radloff, L. S. (1977). The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. http://doi.org/10.1177/014662167700100306

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, *69*(4), 719–727. http://doi.org/10.1037/0022-3514.69.4.719

Schat, A. C., Kelloway, K. E., & Desmarais, S. (2005). The Physical Health Questionnaire (PHQ): Construct validation of a self-report scale of somatic symptoms. *Journal of Occupational Health Psychology*, *10*(4), 363–381. http://doi.org/10.1037/1076-8998.10.4.363

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. http://doi.org/10.1007/s11336-008-9101-0

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 300–308. http://doi.org/10.1037/1040-3590.7.3.300

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The gad-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. http://doi.org/10.1001/archinte.166.10.1092

Stiglitz, J., Sen, A., & Fitoussi, J.-P. (2009). *The measurement of economic performance and social progress revisited* (Documents de Travail de l'OFCE No. 2009-33). Observatoire Francais des Conjonctures Economiques (OFCE). Retrieved from http://econpapers.repec.org/paper/fcedoctra/0933.htm

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment*, *17*(3), 222–232. http://doi.org/10.1027//1015-5759.17.3.222

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*(7524), 550–553. http://doi.org/10.1038/514550a

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and Solutions in Human Assessment* (pp. 41–71). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4615-4397-8_3

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.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. http://doi.org/10.1037/0022-3514.54.6.1063

Wood, A. M., Froh, J. J., & Geraghty, A. W. A. (2010). Gratitude and well-being: a review and theoretical integration. *Clinical Psychology Review*, *30*(7), 890–905. http://doi.org/10.1016/j.cpr.2010.03.005

Wood, A. M., Joseph, S., Lloyd, J., & Atkins, S. (2009). Gratitude influences sleep through the mechanism of pre-sleep cognitions. *Journal of Psychosomatic Research*, *66*(1), 43–48. http://doi.org/10.1016/j.jpsychores.2008.09.002

Wood, A. M., Linley, A. P., Maltby, J., Baliousis, M., & Joseph, S. (2008). The authentic personality: A theoretical and empirical conceptualization and the development of the Authenticity Scale. *Journal of Counseling Psychology*, *55*(3), 385–399. http://doi.org/10.1037/0022-0167.55.3.385

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research. A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. http://doi.org/10.1177/0011000006288127

Table 1: Commonly used and well-validated psychometric measures assessing well-being, mental health, physical health, personality, and related constructs

Scale	Authors and date	Number of items	What it measures
Subjective Well-Being			
Satisfaction with Life Scale (SWLS)	(Diener, Emmons, Larsen, & Griffin, 1985)	5	An individual's global overview of their life. A one-item version of this scale is typically contained within large nationally representative datasets used by economists.
Positive and Negative Affect Schedule (PANAS)	(Watson, Clark, & Tellegen, 1988)	20	Reflects an individual's positive and negative moods at a specific time (e.g., "right now", "last month"). Together with life satisfaction, this can be used to operationalise "Subjective Well-being" (SWB) or the "pleasantness" of an individual's life.
Psychological Well-Being (PWB)	(Ryff & Keyes, 1995)	Variable	Perceptions of; positive relationships, personal growth, autonomy, environmental mastery, purpose in life and self-acceptance.
Happiness Scale	(Lyubomirsky & Lepper, 1999)		Sidesteps the issue of defining happiness by simply asking participants in various ways whether they are happy, without providing a definition.
Mental Health			Health economists are often interested in specific mental health conditions.
General Health Questionnaire (GHQ)	(Goldberg & Williams, 1991)	Variable	Initially designed as a general screening measure to identify participants in large datasets with probable mental health problems. Often included in large nationally representative datasets used by economists.
Centre for Epidemiologic Studies Depression (CES-D) Scale	(Radloff, 1977)	20	One of the most commonly used measures of depression.
Patient Health Questionnaire-9 (PHQ-9)		9	This is the best of the very short depression scales.
Depression, Anxiety, and Stress Scale		21	Measures three common mental health concerns.

(DASS-21)			
Perceived Stress Scale (PSS)	(Cohen, 1988)	10	Assesses stress based on a conception of environmental demands exceeding coping ability.
Generalised Anxiety Disorder-7 (GAD-7)	(Spitzer, Kroenke, Williams, & Löwe, 2006)	7	Focuses on general feelings of anxiety.
Physical Health			Physical health is often better rated objectively, although there are times when subjective reports of health are needed, and these are more common in large datasets.
SF-36	(Ware & Sherbourne, 1992)	36	The most common measure of subjective health, it provides a total score and sub-scores on sub-domains of health.
Physical Health Questionnaire	(Schat, Kelloway, & Desmarais, 2005)	14	Less commonly used, but useful in measuring the milder psychosomatic complaints (sleep disturbance, headaches, gastrointestinal problems, and respiratory infections) so may pick up non-clinical sub-optimal health.
Pittsburgh Sleep Inventory	(Buysse, Reynolds III, Monk, Berman, & Kupfer, 1989)	19	Measures various aspects of impaired sleep.
Personality			Personality is normally assessed via the Big Five.
Big Five Inventory	(John & Srivastava, 1990)	44	One of the most respected measures of the Big Five.
Mini-International Personality Item Pool (Mini-IPIP)	(Donnellan, Oswald, Baird, & Lucas, 2006)	20	A good measure of the Big Five that has the practical minimum of items per factor.
Ten Item Personality Inventory	(Gosling, Rentfrow, & Swann Jr., 2003)	10	The shortest Big Five inventory for use only where the mini-IPIP is not possible.
NEO PI-R	(Costa & McCrae, 2008)	120 to 360	Comprehensively measures the Big Five and the 6 facets of each for a total of 30 traits.
Other measures			
Perceived Social Support (PSS)	(Cohen, Mermelstein, Kamarck, & Hoberman, 1985)		Assesses people's perception of how much practical, emotional, and guidance support they have
Brief Cope	(Carver, 1997)		Assesses the habitual adaptive and maladaptive coping strategies people use to

			deal with stressful situations
Social Desirability Scale-17 (SDS-17)	(Stöber, 2001)	17	Provides a set of plausible socially desirable statements with which everyone would like to agree, but no one can. High scores indicate the person is not answering a questionnaire accurately.
Gratitude Questionnaire-6	(McCullough, Emmons, & Tsang, 2002)	6	Gratitude is one of the strongest predictors of well-being and an emotion that Adam Smith believed to be essential to most successful transactions (see Wood, Froh, & Geraghty, 2010).