Is Construct Validation Valid?

Anna Alexandrova and Daniel M. Haybron*†

What makes a measure of well-being valid? The dominant approach today, construct validation, uses psychometrics to ensure that questionnaires behave in accordance with background knowledge. Our first claim is interpretive—construct validation obeys a coherentist logic that seeks to balance diverse sources of evidence about the construct in question. Our second claim is critical—while in theory this logic is defensible, in practice it does not secure valid measures. We argue that the practice of construct validation in well-being research is theory avoidant, favoring a narrow focus on statistical tests while largely ignoring relevant philosophical considerations.

1. Introduction. What makes a measure of well-being valid? A major project in today's social and medical sciences is measurement of happiness, life satisfaction, and perceived quality of life using self-reports. When questionnaires used to elicit these reports obey the principles of psychometrics, they are considered to be valid measurement tools. Central to this project is *construct validation*—a method for checking the consilience of questionnaires with the background knowledge about the property in question.

In this article we focus on construct validation of measures of self-reported states relevant to well-being. There is perhaps more to well-being than subjective states such as happiness or satisfaction, but we put this concern aside. How an agent feels and judges their life is undoubtedly relevant to their overall well-being—any theorist accepts that much. So evaluating standard measurement tools for detecting these feelings and judgments is important regard-

*To contact the authors, please write to: Anna Alexandrova, Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, UK; e-mail: a.a.alexandrova@gmail.com. Daniel M. Haybron, College of Arts and Sciences, Department of Philosophy, Saint Louis University, Verhaegen Hall, 3634 Lindell Blvd., St. Louis, MO 63108; e-mail: haybrond@slu.edu.

†The authors are equally and jointly responsible for the contents. They thank the anonymous referees, Valerie Tiberius, Colin DeYoung, and Elina Vessonen for valuable comments.

Philosophy of Science, 83 (December 2016) pp. 1098–1109. 0031-8248/2016/8305-0038$10.00
Copyright 2016 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/687941 Published online by Cambridge University Press

less of our philosophical persuasion on the nature of well-being. The real question is whether construct validation evaluates these questionnaires in a fair way.

Our first claim is an explicit statement of the logic of the process, something philosophers have not done so far. Construct validation, we argue, follows a coherentist spirit according to which measures are valid to the extent that they cohere with theoretical and empirical knowledge about the states being measured. In theory this is a defensible approach to measurement, but in practice the current procedures of validation do not respect all sources of knowledge about well-being, and this is our second claim. Construct validation is in fact dangerously *theory avoidant*, failing to respect a core commitment of any plausible theory of well-being, namely, that well-being is a normative category. This constraint implies that measures of subjective states relevant to well-being need to be judged on their normative validity in addition to other characteristics. The current almost exclusive attention to the statistical correlations between questionnaires and questionnaire items does not provide sufficient constraints to weed out weak measures. We close with a suggestion for how construct validation can be improved.

2. What Is Construct Validation? The first order of business is to get clear on the logic behind the procedure. The psychometric tradition in the social sciences has historically specialized in tests and questionnaires for detecting unobservable attributes such as intelligence and personality traits. Today for virtually all researchers who wish to measure any attribute on the basis of self-reports or performances in tests, psychometric validation remains the obligatory procedure. The practitioners of the new science of well-being—psychologists, sociologists, clinical scientists—have also embraced questionnaires and, with that, psychometric validation.

Questionnaires used in well-being research range from gauging a person's feeling ("How anxious do you feel?") to gauging their judgments ("Is your life going well according to your priorities?") to gauging their perception of facts deemed important ("Do you feel in control of your circumstances?"). They can be longer or shorter and administered through various media. Some well-known questionnaires include the Satisfaction with Life Scale (SWLS; Diener et al. 1985), the Positive and Negative Affect Scale (PANAS; Watson, Clark, and Tellegen 1988), and the Nottingham Health Profile (Hunt et al. 1981), which measure life satisfaction, happiness, and health-related quality of life, respectively.

Validation of these scales follows a typical pattern described in measurement textbooks and articles on validation (Simms 2008; de Vet et al. 2011). First, researchers define the construct to be measured by elaborating its scope and limits. This is the conceptual stage in which the meaning of the concepts in question is discussed, invoking anything from philosophical

https://doi.org/10.1086/687941 Published online by Cambridge University Press

theories to untutored intuitions to dictionary definitions. For example, the scope of happiness is often deemed to be positive and negative affect, while the scope of satisfaction with life is deemed a cognitive judgment about one's conditions and goals. In the second stage, researchers choose a measurement method (a questionnaire, a test, or a task), select the items (what questions? what tasks?), and settle on the scoring method. In the third and final stage, the instrument is tested for its validity. We focus on this last step, because it is supposed to discipline all the free philosophizing that happens in the earlier stages with the hard tools of psychometrics. What are those tools?

It is hard to speak of a psychometric method in general because the methods are numerous and constantly evolving.[1] But in the case of well-being measures, validation frequently involves *factor analysis*: when hundreds of subjects fill out the same questionnaire, perhaps several times over a period, it is possible to observe the correlations between responses to different items. These correlations are then used to show that there are one or more clusters of items called 'factors' that account for the total information. Scientists speak of factor analysis as extracting "a manageable number of latent dimensions that explain the covariation among a larger set of manifest variables" (Simms 2008, 421).[2] Explanation is here used in an entirely phenomenological sense as saving the phenomena (the phenomena being the total data generated by administering the questionnaire in question), rather than stating the causes of the phenomena. For example, the SWLS is a popular five-item Likert scale for measuring the cognitive aspect of subjective well-being, that is, the extent to which subjects judge their life to be satisfactory. Factor analysis identified all five items to be measuring the same latent variable because a single factor accounted for 66% of the variance in the data (Diener et al. 1985). Other scales may turn out to gauge more than one dimension.

The next step of the testing stage is to check that the behavior of these factors accords with other things scientists know about the object in question. In the case of subjective well-being, this knowledge includes how people evaluate their lives and surroundings, what behavior results from these evaluations, and what other people who know the subjects say about them. For example, the aforementioned SWLS, according to its authors, earned

1. Sawilowsky (2007) summarizes the state of the art.

2. There is a difference between *exploratory* and *confirmatory* factor analysis (see de Vet et al. 2011, 169–72, among other places). The former is used to reduce the number of items in a questionnaire by identifying the one(s) that best predict the overall ratings. The latter, on the other hand, tests that the factors that best summarize the data also conform with a theory of the underlying phenomenon if there is one. This distinction is not important for the present argument.

https://doi.org/10.1086/687941 Published online by Cambridge University Press

construct validity when Diener and his colleagues compared responses on the SWLS to responses on other existing measures of subjective well-being and related constructs such as affect intensity, happiness, and domain satisfaction. The findings confirmed their expectation that SWLS scores correlate highly with those measures that also elicit a judgment on subjective well-being and less so with measures that focus only on affect or self-esteem or other related but distinct notions. One piece of evidence in favor of SWLS was that the scores of 53 elderly people from Illinois correlated well to the ratings this same population received in an extended interview about "the extent to which they remained active and were oriented toward self-directed learning" (Diener et al. 1985, 73). How strong was the correlation? It was $r = 0.43$, which is adequate by the standards of the discipline. Since 1985, SWLS has continued to be scrutinized for its agreement with the growing data about subjective well-being. Individual judgments of life satisfaction have been checked against the reports of informants close to the subjects (Schneider and Schimmack 2009). Proponents of SWLS argue that it exhibits a plausible relationship with money, relationships, suicide, and satisfaction with various domains of life, such as work and living conditions.[3]

Now we are in a position to formulate a logic for psychometric validation that we believe captures these practices:

> *Implicit Logic.* A measure *M* of a construct *C* is validated to the extent that *M* behaves in a way that respects three sources of evidence:
>
> 1. *M* is inspired by a plausible theory of *C* specified in stage 1.
> 2. Subjects reveal *M* to track *C* through their questionnaire answering behavior.
> 3. Other knowledge about *C* is consistent with variations in values of *M* across contexts.

The first condition captures the role of philosophizing about the nature of *C* in the first stage of measure development. There are no strong criteria for what makes a conception of *C* plausible and how elaborate it should be. The second condition specifies the assumption behind factor analysis.[4] The third acknowledges that scientists go beyond the merely internal analysis of the scale: a valid measure correlates with indicators that our background knowledge says it should and does not correlate with indicators that it shouldn't. To-

3. See Diener et al. (2008, 74–93) for summary and references.

4. For convenience we are focusing on the practice of factor analysis, even though not all construct validation procedures that concern us involve it—e.g., the validation of single-item measures.

https://doi.org/10.1086/687941 Published online by Cambridge University Press

gether the three conditions capture what it takes for a measure to be declared valid,[5] but they do not explain the reasons why this inference works. So the next step is to evaluate Implicit Logic.

3. Construct Validation Is Good in Theory. Construct validation as described above conceives of measurement as part of theory development and validation as part of theory testing. On the original proposal formulated in the classic 1955 article by Lee Cronbach and Paul Meehl, construct validation consists in testing the nomological network of hypotheses in the neighborhood of the construct in question (Cronbach and Meehl 1955). To measure x, we need to know how x behaves in relation to other properties and processes that are systematically connected with x by lawlike regularities. Something like this view is still the consensus: "To determine whether a measure is useful, one must conduct empirical tests that examine whether the measure behaves as would be expected given the theory of the underlying construct" (Diener et al. 2008, 67).

We believe that this vision of measure validation is defensible. Its spirit is remarkably similar to the coherentist vision that characterizes recent work on measurement of physical quantities (Chang 2004; van Fraassen 2008; Tal 2013). These philosophers emphasize that the outlines of the concept in question, be it temperature or time, and the procedure for detecting it are settled not separately but iteratively, checking and correcting one against another. Similarly in our case, the initial philosophical judgment about the nature of happiness or quality of life is coordinated with other constraints such as the statistical features of the questionnaires and the background knowledge about behavior, related indicators, and ratings of informants. The resulting measurement tools can be deemed valid to the extent that they accommodate all evidence.

The above vision appears to contrast with cases where measurement starts with a set of observable relations (e.g., rigid rods of different lengths, or choices of different goods by an agent) and proceeds via axioms to numerical structures (such as a sequence of real numbers to represent length or utility function). The latter picture is often associated with the *representational theory of measurement*. According to this, a measure is valid if there is a demonstrated homomorphism between an observable relation and a numerical relational structure (Krantz et al. 1971). The economic approach to welfare measurement via gross domestic product and other economic indicators seems to follow this logic because it relies, in part, on axioms that

5. There are, of course, other kinds of validity. We concentrate on construct because among measurement theorists the consensus seems to be that construct validity encompasses all other types of validity, such as criterion, predictive, discriminant, and content validity (Strauss and Smith 2009).

https://doi.org/10.1086/687941 Published online by Cambridge University Press

relate preferences to utility. Some commentators conclude that since the psychometric approach does not rely on axioms, it is therefore not in keeping with the representational theory (Angner 2009).

We make no such claims. It may well be that the psychometric approach is not a tradition of its own and that it too needs something that has played the role of axioms in the representational theory.[6] Perhaps step 1 of our Implicit Logic aims at this goal by delineating the bounds of the concept in question. All we claim is that the ideal behind construct validation is to formulate reliable scales that accord with background knowledge. If this process works, it should be enough for measurement. But does it?

4. Construct Validation in Practice. Things look worse in practice than in theory. Although questionnaires are validated against a broad range of evidence, psychometricians are selective about what counts as evidence in favor of or against construct validity. We see two problems that illustrate this selectivity. First, the existing data used to validate questionnaires do not provide sufficient constraints to weed out the poor ones. Secondly, a legitimate source of evidence about the nature of states relevant to well-being—philosophical theorizing—is either never used or else overridden by statistical considerations. These are the two senses in which construct validation is *theory avoidant*, sacrificing valid theoretical knowledge for statistics for no good reason.

As step 3 of our Implicit Logic shows, researchers base judgments of validity mainly on whether the measure in question exhibits plausible-seeming correlations with relevant-seeming variables. This is not unreasonable, since correlational data are the main source of empirical evidence at hand, and there is something of a chicken-and-egg problem in that, if we already knew exactly what correlations a measure should exhibit, we might not have much need for the measure. One piece of evidence that a well-being measure is valid, for instance, might be that it correlates to some significant degree with money. But then, on the other hand, the correlation between well-being and money may be precisely one of the things we hope to find out using the measure. Psychometricians have their work cut out for them.

It makes sense, then, that validation procedures should be flexible and holistic: we see whether, on balance, the measure behaves in a way that makes sense. While correlations with any given variable might prove to be surprising, the overall pattern of correlations should not, in general, be too much of a surprise. When we do get broadly unexpected results—as might have been the case, for instance, when research seemed to indicate that happiness was so strongly prone to adaptation as to be nearly immuta-

6. See Cartwright and Bradburn (2011) on the importance of representation in social measurement, where concepts are often fuzzy and multitudinous.

https://doi.org/10.1086/687941 Published online by Cambridge University Press

ble (Lykken and Tellegen 1996)—then either we need some theoretical framework to make sense of it, say, that happiness is strongly governed by homeostatic mechanisms that keep the individual hovering around a given "set point," or we should suspect that the measures are not in fact valid (or that the results are otherwise spurious).

The trouble is that what counts as a "plausible correlation" is a rather elastic quantity, both vague and open to the interpretive predilections of the investigator, whose judgment in the matter may be less than impartial. The problem is particularly acute in well-being research, where it can seem as if nearly everything correlates substantially with nearly everything else. Moreover, commonsense views of well-being tend to be both expansive and incoherent; it is only somewhat exaggerated to say that just about anything one might care to venture about well-being—money buys happiness, money doesn't buy happiness—is already part of the folklore.

Take a long list of variables that seem like they might be related to well-being—money, relationships, health, education, work, and so on. Imagine two measures, A and B, each of which correlates substantially with nearly all of these variables, while also differing greatly in what those correlations are. One suggests that relationships are more strongly related to well-being than money, while the other has the reverse implication, and so forth. It seems entirely possible that both measures could reasonably be deemed to exhibit "plausible correlations" and generally pass as valid measures of well-being. It is also possible that one of those measures is in fact valid, while the other is not: A gets the correlations essentially right, while B gets them wrong.

This sort of scenario is not merely a theoretical possibility. Recent studies have found that life evaluation and affect measures of well-being give importantly different results, and some researchers have taken the differences to indicate that life evaluation metrics (such as the SWLS) are superior on the grounds that they are, or are claimed to be, more sensitive to life circumstances—generally, correlating more strongly with quantities that have traditionally interested policymakers such as income, governance, freedom, and so on (Helliwell, Layard, and Sachs 2012). One question, to which we will return shortly, is whether the affect measures in question are themselves well designed. More pertinent for current purposes is this: why should we assume that the better measure must correlate more strongly with those variables? Suppose that hedonism, one of the main theories of well-being in the literature, is in fact correct. In that case, perhaps the best interpretation of the data is that well-being isn't very sensitive to life circumstances. (Of course, those variables, like good governance, might matter a great deal for reasons of justice, or some other reason.)

Alternatively, perhaps the "life circumstances" on which these researchers are focusing just aren't the ones that matter most for well-being. An im-

https://doi.org/10.1086/687941 Published online by Cambridge University Press

portant article discussing data from the same global survey, for example, reports that while life evaluation metrics do indeed track "material prosperity" more strongly, the affect measures better correlate with what the authors call "psychosocial prosperity": whether people reported being treated with respect in the previous day, had friends to count on, learned something new, did what they do best, or chose how their time was spent (Diener et al. 2010). It would not be eccentric to suggest that these are just the sorts of variables that seem most obviously to matter for well-being, and to which good measures of well-being ought to be sensitive. Perhaps, then, it is the affect measure, and not life evaluation, that offers a more meaningful picture of well-being.

Or perhaps not. Our point is not to endorse or critique either sort of measure.[7] There may be other reasons to favor life evaluation measures, and there are differences in the data sets being used by these investigators that we cannot assess here. The point is just to illustrate how two prominent measures could both be deemed valid measures of well-being by prevailing standards, though they have very different statistical properties—and, crucially, statistical tests alone cannot tell us which is the superior instrument. We need to appeal to theoretical considerations as well: what conception of well-being is relevant here? Given our best understanding of human well-being, what sorts of factors should a good measure correlate most strongly with? Is the measure that more closely tracks money and stuff likely to be a better indicator of well-being than one that tracks relationships and meaningful work? If we do not take these theoretical questions seriously, ideally before testing our instruments, we risk settling on whatever measures are most convenient, most congenial to our personal views, or simply ours, and not someone else's.

One form of theory avoidance, then, can lead us to focus on the wrong correlations, or have the wrong ideas about what the right correlations are: the statistical data alone do not provide sufficient constraints to allow us to assess the validity of a measure. In a second form, theory avoidance can have us measuring the wrong variables altogether, because our instruments are insufficiently grounded in theoretical considerations that might provide a rationale for their design. We illustrate with an example of a popular affect questionnaire known as PANAS. PANAS assesses the relative prevalence of positive over negative mood and is commonly used to measure the affective dimensions of subjective well-being. This 20-item questionnaire

7. We may be seeming to mix apples and oranges here, as life evaluation and affect measures aren't even supposed to be measures of the same construct. In fact, however, this is not entirely true: while their proximal concerns are quite distinct, both are often posited and deployed more fundamentally as general metrics of well-being, aimed at giving a rough snapshot of overall welfare.

https://doi.org/10.1086/687941 Published online by Cambridge University Press

asks subjects to rate themselves on whether they feel enthusiastic, interested, excited, strong, alert, proud, active, determined, attentive, inspired, and so on (Watson et al. 1988). All these items have passed factor analysis and other standard psychometric tests. But note that absent from this list are cheerfulness, joy, laughter, sadness, depression, tranquillity, anxiety, stress, weariness—emotions that are intuitively far more central to a happy psychological state and to well-being. This is because the authors of PANAS arrived at the list of items by testing a long list of English mood terms and paring it down via factor analysis, so that a longer list would not yield appreciably different results.

Such a procedure allows investigators to avoid hard theoretical questions about which taxonomy of emotional states to employ, or which states are most relevant to well-being. But for the same reason, there is little reason to expect such a method to yield a sound measure of well-being, or even of emotional well-being. Rather, what is being assessed, roughly, is the number of English mood terms that apply to the respondent—or rather, the number of terms from a list of words that survived factor analysis. But, first, this leaves the measure prey to the vagaries of common English usage and folk psychology—potentially important emotional phenomena may not be prominent in the vocabulary of a given language, or may not be correctly classified as emotional, and so may be omitted from the measure. Of particular concern here are relatively diffuse background states—anxiety, stress, peace of mind (not on the list)—that are quite important for well-being yet easily overlooked, resulting in a kind of "streetlight" problem where we end up looking where the light is best, rather than where the keys are.

Second, some states are presumably more important for well-being than others; feelings of serenity or joy (not on the list) probably count for more than feeling "attentive" or "alert" (on the list), and indeed some of the PANAS items might barely deserve inclusion at all, if our interest is in assessing well-being. Yet a term like "attentive" might exhibit quite distinctive correlations and thus make it on the list, while other more salient terms are left by the wayside.

The worries here essentially amount to saying that you can't get the right measure without attending to theoretical considerations—namely, what do our best theories tell us are the emotional states that might matter for well-being? For example, one of the authors recently proposed an account of emotional well-being, or happiness, that divides emotional states into three broad types—representing functional responses to different types of well-being-relevant information regarding matters of security, opportunity, and success—and further posits emotional well-being as a central element in an account of well-being (Haybron 2008, 2013). Whether or not that taxonomy is the right one to employ in well-being measures, some such account

https://doi.org/10.1086/687941 Published online by Cambridge University Press

could provide a theoretically motivated basis for developing affect-based well-being instruments.

We do not deny that PANAS is useful or exhibits some desirable statistical properties, and perhaps it does provide a reasonable, if somewhat opaque, metric of well-being. As before, our purpose is not to critique a particular measure so much as to illustrate how practices of construct validation can be seriously inadequate given the ease with which they can fail to attend seriously to theoretical concerns. While we have not tried to document the extent of the problem and have focused mainly on illustrating the risks, that there is some problem should be uncontroversial. The risks, we think, are not infrequently realized if only because the examples discussed here, the SWLS and PANAS, are very popular. The problem here resembles a complaint often lodged against philosophers' conceptual or linguistic analyses, namely, a heavy reliance on the investigators' hunches or intuitions, without adequate attention to the theoretical motivation, or lack thereof, for reaching a certain view. This is not just a hazard for philosophers.

5. What Is to Be Done? It is understandable that social scientists, like other researchers, will want to focus their efforts where their competence and interests are greatest. The theory-avoidant status quo has developed in psychology owing to its operationalist heritage, which was key to its establishment as a 'hard' science; even today, psychologists insist that although building substantive theories of subjective well-being is a worthy enterprise, they are not trained to do so and it is safer to tread close to the easily observable and reproducible results of psychometrics. Any proposal for reform should respect the fact that this status quo is unlikely to change in any deep ways. But correlation mongering is no substitute for theory, and so we urge that construct validation protocols also assess the *normative validity* of measures. The normative validity of a measure of, say, happiness, is the extent to which this measure respects the importance of happiness for well-being, since well-being is the ultimate object of concern for the scientific project in question. We conceive of normative validity as a fourth condition on Implicit Logic in addition to the three existing ones: a measure *M* must respect what is important about construct *C*. Just as philosophers relying on empirical assumptions are increasingly expected to engage with the relevant scientific literatures, so too should empirical researchers attend to the literature that bears on the key philosophical assumptions they are making.

We are under no illusions that this is a lot to ask of scientists whose identity often enough consists in not being philosophers. Besides, the very question of normative validity can be a genuinely difficult one—philosophers do disagree about the importance of, say, life satisfaction for well-being. Nevertheless, a scholarly convention to discuss normative validity at least briefly in articles on validation would go some way toward flagging this issue. At

https://doi.org/10.1086/687941 Published online by Cambridge University Press

the very least, if there is no theory of well-being according to which the construct in question is important, that should count against a measure.

The science of well-being makes no pretense of being value-free in one clear sense: well-being is a value worth understanding and pursuing. The eager and successful policy engagement of the prominent figures in this field attests to this therapeutic mission. From this point of view our proposal is quite tame—we merely try to show how the measurement and validation practices of the science of well-being can catch up to the already-existing normative ambition.

REFERENCES

Angner, Erik. 2009. "Subjective Measures of Well-Being: Philosophical Perspectives." In *The Oxford Handbook of Philosophy of Economics*, ed. Harold Kincaid and Don Ross, 560–79. Oxford: Oxford University Press.

Cartwright, Nancy, and Norman Bradburn. 2011. "A Theory of Measurement." In *The Importance of Common Metrics for Advancing Social Science Theory and Research: Proceedings of the National Research Council Committee on Common Metrics*, 53–70. Washington, DC: National Academies.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4): 281–302.

de Vet, Henrica C. W., Caroline B. Terwee, Lidwine B. Mokkink, and Dirk L. Knol. 2011. *Measurement in Medicine: A Practical Guide*. Cambridge: Cambridge University Press.

Diener, Ed, Robert A. Emmons, Randy J. Larsen, and Sharon Griffin. 1985. "The Satisfaction with Life Scale." *Journal of Personality Assessment* 49 (1): 71–75.

Diener, Ed, Richard E. Lucas, Ulrich Schimmack, and John Helliwell. 2008. *Well-Being for Public Policy*. New York: Oxford University Press.

Diener, Ed, Weiting Ng, James Harter, and Raksha Arora. 2010. "Wealth and Happiness across the World: Material Prosperity Predicts Life Evaluation, Whereas Psychosocial Prosperity Predicts Positive Feeling." *Journal of Personality and Social Psychology* 99 (1): 52.

Haybron, Daniel. M. 2008. *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being*. New York: Oxford University Press.

———. 2013. *Happiness: A Very Short Introduction*. New York: Oxford University Press.

Helliwell, John, Richard Layard, and Jeffrey Sachs. 2012. *World Happiness Report*. Columbia University: Earth Institute.

Hunt, Sonja M., S. P. McKenna, J. McEwen, Jan Williams, and Evelyn Papp. 1981. "The Nottingham Health Profile: Subjective Health Status and Medical Consultations." *Social Science and Medicine. Part A: Medical Psychology and Medical Sociology* 15 (3): 221–29.

Krantz, David, Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. *Foundations of Measurement*. Vol. 1, *Additive and Polynomial Representations*. New York: Academic.

Lykken, David, and Auke Tellegen. 1996. "Happiness Is a Stochastic Phenomenon." *Psychological Science* 7 (3): 186–89.

Sawilowsky, Shlomo. 2007. "Construct Validity." In *Encyclopedia of Measurement and Statistics*, ed. Neil J. Salkind and K. Rasmussen, 179–82. Thousand Oaks, CA: Sage.

Schneider, Leann, and Ulrich Schimmack. 2009. "Self-Informant Agreement in Well-Being Ratings: A Meta-analysis." *Social Indicators Research* 94 (3): 363–76.

Simms, Leonard J. 2008. "Classical and Modern Methods of Psychological Scale Construction." *Social and Personality Psychology Compass* 2 (1): 414–33.

Strauss, Milton E., and Gregory T. Smith. 2009. "Construct Validity: Advances in Theory and Methodology." *Annual Review of Clinical Psychology* 5:1–25.

https://doi.org/10.1086/687941 Published online by Cambridge University Press

Tal, Eran. 2013. "Old and New Problems in Philosophy of Measurement." *Philosophy Compass* 8 (12): 1159–73.

van Fraassen, Bas. C. 2008. *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

Watson, David, Lee A. Clark, and Auke Tellegen. 1988. "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales." *Journal of Personality and Social Psychology* 54 (6): 1063.

https://doi.org/10.1086/687941 Published online by Cambridge University Press