

Self-Informant Agreement in Well-Being Ratings: A Meta-Analysis

Leann Schneider · Ulrich Schimmack

Accepted: 5 January 2009
© Springer Science+Business Media B.V. 2009

Abstract A meta-analysis of published studies that reported correlations between self-ratings and informant ratings of well-being (life-satisfaction, happiness, positive affect, negative affect) was performed. The average self-informant correlation based on 44 independent samples and 81 correlations for a total of 8,897 participants was $r = 0.42$ [99% credibility interval = 0.39|0.45]. Statistically reliable moderators of agreement were construct (life-satisfaction = happiness > positive affect > negative affect), age of the target participant (older > younger), number of informants (multiple > single), and number of items in the measure (multiple > single). The implications for the validity of self-ratings of well-being as indicators of well-being are discussed.

Keywords Well being · Positive affect · Negative affect · Life satisfaction · Happiness · Self-rating · Informant rating · Validity · Measurement

1 Introduction

A major goal of well-being science is to conduct empirical studies of the causal determinants of well-being (Schimmack 2008). To achieve this goal, well-being science—like any other empirical science—needs valid measures of well-being. Despite a large amount of empirical studies of well-being, the validity of well-being measures in these studies remains unknown because few studies define well-being. Without an explicit definition of well-being, it is impossible to determine whether a measure is a valid measure of well-being. That is, in order to determine whether a measure measures well-being, it is necessary to define well-being.

One solution to this problem has been to propose a list of indicators that are likely to correlate with well-being. Although lists vary across studies, a large number of empirical

L. Schneider · U. Schimmack (✉)
University of Toronto Mississauga, Mississauga, ON, Canada
e-mail: uli.schimmack@utoronto.ca

L. Schneider
e-mail: leann.schneider@utoronto.ca

studies have used measures of happiness, life satisfaction, (high) positive affect, and (low) negative affect as indicators of well-being (Diener 1984). If we assume that well-being is reasonably well captured by these indicators, it becomes possible to examine the validity of well-being measures in terms of their ability to measure happiness, life-satisfaction, positive affect, or negative affect. That is, it is possible to examine whether a life-satisfaction measure actually measures life-satisfaction and whether a positive affect measure actually measures variation in positive affect.

As objective measures of happiness, life-satisfaction, positive affect, or negative affect are lacking, it is impossible to examine the validity of these measures directly against an objective criterion. For this reason researchers have to rely on indirect evidence to examine the validity of well-being measures. The indirect approach may explain the contradictory claims about the validity of well-being measures in the literature. Experimental studies often show that it is possible to alter life-satisfaction judgments by manipulating respondents' attention to specific information or mood states. This evidence has led some researchers to propose that well-being measures are invalid (Schwarz and Strack 1999). However, several articles have pointed out the flaws in this line of reasoning (Eid and Diener 2004; Schimmack and Oishi 2005). First, the argument confuses statistical significance with practical significance. The effect size of the experimental manipulations on well-being measures is often rather small. Thus, it is possible that most of the variance in well-being judgments is valid despite statistically significant effects of experimental manipulations. Second, experimental studies may exaggerate effect sizes under naturalistic conditions because experimenters create situations that maximize effects. For example, whereas well-being reports can be influenced by the presence of a disabled confederate in an experiment, the effect size in a naturalistic study would be much smaller because few respondents report well-being in the presence of an individual with a disability (Schwarz and Strack 1999). Moreover, the presence of a well-known disabled family member may not produce the same effects as the presence of a disabled stranger in a laboratory study. Third, many of the influences that produce systematic effects in a controlled experiment produce unsystematic effects under naturalistic conditions. Indeed, Schwarz and Strack (1999) point to low test–retest correlations within a one-hour retest interval as evidence against the usefulness of well-being measures. This argument confuses reliability and validity. Low reliability can easily be increased by using multiple indicators or repeated measurements. Furthermore, random measurement error only attenuates effect sizes, but it does not systematically distort the results of empirical studies. For example, random error can lower the correlation between income and well-being from $r = 0.3$ to 0.2, but it cannot turn a positive correlation into a negative correlation. In short, the demonstration that experimental manipulations have small to moderate systematic effects on well-being reports tells us very little about the validity of well-being measures in naturalistic studies that examine the causes of well-being.

Half a century ago, personality psychologists proposed a procedure that could be used to estimate the amount of valid variance in a measure even if no perfect validation criterion is available (Campbell and Fiske 1959). The indirect assessment of validity requires that a construct (e.g., life satisfaction) has to be measured with at least two methods. If two measures of the same construct are available, the correlation between the two measures provides information about the validity of the two measures. A positive correlation between the two measures demonstrates convergent validity of the two measures. Although two methods are sufficient, the strength of the conclusions increases with the number of independent methods that are used to measure the same construct.

A fundamental assumption of studies of convergent validity is that the true variance in the construct under investigation, for example life-satisfaction, is the only causal factor that

produces the correlation between two measures. This assumption is unlikely to be true if both measures are based on self-ratings. For example, the correlation between self-ratings on Cantril's ladder and the Satisfaction With Life Scale could be influenced by shared memory biases or a tendency to respond in a socially desirable manner. Thus, the main challenge for validation studies of well-being measures is to find multiple measures of well-being that use independent methods. Moreover each method must have at least some validity. If one method has zero validity, it cannot reveal validity in the other measure.

The most widely used approach to obtain multiple independent measures of well-being is to obtain well-being ratings from multiple raters, typically the target of the study (self-rating) and one acquaintance of the target (informant ratings). The correlation between self-ratings and informant ratings reflects the amount of convergent validity in the two measures. Positive correlations between self-ratings and informant ratings of well-being provide some of the strongest evidence for the validity of well-being measures (Diener et al. 1995). However, while self-informant correlations >0 indicate that both measures have some validity, it remains unclear how much of the variance in each well-being measures is valid. In other words, existing evidence is strong enough to reject the nil-hypothesis that well-being measures have zero validity (Cohen 1994). However, the more important question is how valid well-being measures are. The main aim of this article is to estimate the amount of valid variance in well-being measures based on a meta-analysis of self-informant correlations. The second aim of this study is to examine potential moderators of self-informant agreement. This information can be used to develop better measures of well-being.

2 Past Research

Before we examine self-informant agreement of well-being measures, it is instructive to review self-informant agreement for other constructs to obtain a benchmark for self-informant agreement of well-being measures. Numerous studies have examined self-informant agreement for ratings of personality traits like extraversion or conscientiousness. A meta-analysis produced an average agreement of $r = 0.36$ across studies and different personality traits (Connolly et al. 2007). The estimate for studies with close informants was slightly higher, $r = 0.43$. Similar results have been obtained for well-being measures. An early study by Hartmann (1934) produced a correlation of $r = 0.34$ between self-ratings and average informant ratings by four informants in a sample of 195 students. Adjusted for the retest reliability of $r = 0.70$ reported in the article, the correlation would be $r = 0.41$ without random measurement error. Even with the relatively large sample size of 195 participants the 99% confidence interval for this estimate ranges from 0.24 to 0.58, indicating that the true correlation could be as low as $r = 0.24$. Moreover, the study does not provide information about the amount of valid variance in self-ratings. A correlation of 0.24 could be the result of perfect validity in the self-ratings and relatively low validity of informant ratings ($1 \times 0.24 = 0.24$) or vice versa (0.24×1). In the latter scenario, the amount of valid variance in the self-rating of well-being would be only 6% corrected for unreliability and 4% in the observed single-item measure of well-being. The true validity is likely to be higher, but it is not very reassuring to realize that only a small percentage of the variance in self-ratings of well-being may be valid. About 30 years later, this study remained the only noteworthy study of convergent validity of well-being measures (Wilson 1967). The evidence base increased over the following 20 years. A review article that focused on the Satisfaction With Life Scale reported self-informant correlations of six independent studies (Pavot and Diener 1993a). The unweighted average correlation was $r = 0.44$. Adjusting this correlation for

unreliability in the self-ratings yields a correlation of $r = 0.48$, which is similar to the results in Hartmann's (1934) seminal study, and to self-informant agreement for ratings of personality traits. Thus, we predicted that our meta-analysis would reveal an average self-informant agreement for well-being ratings of around $r = 0.4$.

3 Moderator Variables

Previous studies of self-informant agreement in well-being ratings have been limited to demonstration of convergent validity (i.e., rejection of the nil-hypothesis). However, an important question is whether self-informant correlations systematically vary as a function of other factors. A few factors are self-evident. First, the number of items in a measure will influence agreement for the simple reason that scales with more items are more reliable and reliability influences validity. Second, many studies averaged across reports by multiple informants. Number of informants also moderates correlations because averages of multiple raters are more reliable than ratings by a single rater. In addition, averages reduce systematic error variance that is unique to a single rater, which increases the amount of valid variance in the aggregated measure. A more substantive finding is that the validity of informant ratings increases with closeness to the target (Connolly et al. 2007).

Another potential moderator is visibility, that is, the availability of observable cues (e.g., non-verbal behavior) that provide valid information about the characteristic that is being judged. As a result, self-informant agreement tends to be higher for extraversion than neuroticism (Connolly et al. 2007). As extraversion is more highly related to positive affect than negative affect, we predict higher self-informant agreement for positive affect than negative affect. Another reason for this prediction would be that people are more likely to hide negative feelings from others. Based on the visibility hypothesis, we might expect the lowest agreement for global judgments of well-being because these judgments do not have overt behavioral correlates. However, informant ratings do not have to rely on mere behaviors to be valid. Another way to judge others well-being is to base it on one's knowledge of the targets' ideals and knowledge about the target's actual life-circumstances. Finally, informant ratings may also be based on verbal information. In this regard, it is more informative if somebody tells us they hate their job than non-verbal cues of positive affect in a few situations. Thus, while higher visibility of behavior would favor affective measures of well-being, other sources of self-informant agreement may favor global judgments of well-being. For this reason, we did not make an a priori prediction about the difference between affective and cognitive measures of well-being.

Another potential moderator of self-informant agreement could be age. The well-being of younger individuals may be more variable, which makes it more difficult to judge. Moreover, younger respondents may have less stable perceptions of their well-being, which would lower validity in self-ratings and in turn, convergent validity.

4 Method

4.1 Literature Search

We searched for articles with self-informant correlations using the web of science database. First, a keyword search was done with various combinations of the following terms: positive affect, negative affect, life satisfaction, happiness, (subjective) well-being, SWB,

informant(s), self-informant, agreement, multimethod, convergent, convergence, correspondence, concordance, proxy, proxies, and collateral. The retrieved articles were scanned for relevance, and added to our collection if they fit our criteria. Secondly, we checked the references cited by relevant articles for further articles.

4.1.1 Inclusion Criteria

Articles were only included in our analyses if they reported a Pearson correlation between a self and informant rating of positive affect, negative affect, life satisfaction or happiness. Studies that examined ratings of momentary affect (how do you feel right now) were excluded because a single rating of momentary affect is not a valid measure of well-being. If a study included self-informant correlations for targets with intellectual disabilities, it was not added to our analyses. We limited our analysis to studies that used friends, family, or a combination of both as informants mainly because few studies used other raters as informants (e.g., teachers, co-workers). As a result, the range of closeness is relatively small and it may be difficult to find evidence for effects of closeness on self-informant agreement. However, we predicted higher correlations for family members as informants based on findings for personality traits (Connolly et al. 2007).

4.2 Study Variables

4.2.1 Age

When the mean age was reported for the sample of participants from which the self-informant correlation was given, we recorded it. In some cases the mean age for the total sample was reported (including participants with no informant information) and we used this age to represent the mean age for the subsample that we used. For many articles, no age was given but it was reported that data was from a student sample. In this case we assigned a mean age of 20 for the participants in these samples. For the study by Watson and Humrichouse (2006), no mean age was given for participants at Time 2 data collection, but it was calculated using the mean time elapsed from Time 1, and the mean age of participants at Time 1. Due to the uneven distribution of participants' ages across studies, we used age groups in our moderator analysis. We assigned target participants under the age of 24 to the younger group, and target participants over the age of 24 to the older group.

4.2.2 Informant Type

For each study, the type of informant was categorized as Friend (0) or Family (1). Dating couples were counted as a friend relationship. When there was a mix of both family and friends who made ratings, or when there was only one informant, but in the sample there was a mix of informants who were friends or family, then the study was included in a mixed Family/Friend category (0.5). Based on our coding scheme, a positive moderator effect reveals higher self-informant agreement for family members than friends.

4.2.3 Number of Informants

The number of informants (or mean number, if a mean was given and the number of informants was not the same for each participant) was recorded for each study. For studies

where there was a variation in the number of informants but the mean was not given, the lowest number of informants was used.

4.2.4 Number of Items

Most studies used the same measure for self-ratings and informant ratings. Thus, only one variable was used to distinguish studies. Fifty-six correlations were based on multiple-item scales (coded 1), 13 used single-item scales (coded 0), and 12 used single-item scales for one type of rating and multiple-items for the other type of rating (coded 0.5).

5 Results

Table 1 lists the 81 self-informant correlations based on 44 independent samples ($N = 8,897$) uncovered by the literature search. The data were analyzed using weighted mixed-model regression analysis, implemented in MPLUS5 (Muthén and Muthén 2007). A mixed model was used because some articles contributed more than one self-informant correlation based on the same sample. As a result, self-informant correlations were clustered within samples. A weighted model was used because larger samples have smaller standard errors. However, an unweighted analysis produced very similar results.

A null model (without predictor variables) was used to estimate the weighted average correlation and its 99% credibility interval (Whitener 1990). The correlation was $r = 0.42$ [99% credibility interval = 0.39|0.45]. This finding is quite similar to the agreement for judgments of personality traits (Connolly et al. 2007), indicating that well-being is neither easier nor harder to judge than personality traits. The null model also revealed significant amount of variance across samples, $V = 0.009$, $SE = 0.002$, $SD = 0.095$, and within samples across different measures, $V = 0.005$, $SE = 0.002$, $SD = 0.071$. The next model added predictor variables to examine the contribution of moderator variables to this variation.

Due to the low power of these analyses, we used a fixed effects model and a 95% confidence interval to test statistical significance of moderator effects. Numerous significant moderator variables were found (Table 2). First, global judgments produced higher correlations ($r = 0.35$) than positive affect ($r = 0.24$), which produced higher correlations than negative affect ($r = 0.18$). There was no reliable difference between life-satisfaction or happiness ratings. Correlations were higher for older participants ($r = 0.32$) than for younger participants ($r = 0.24$). Correlations were higher for studies with multiple informants ($r = 0.31$) than for studies with a single informant ($r = 0.24$), but type of informant did not influence correlations notably. Studies with multiple item scales produced higher correlations ($r = 0.34$) than studies with single-item self-ratings ($r = 0.24$). Publication year was not a significant moderator.

The moderator analysis would imply that the maximum agreement could be obtained in a study with a multiple-item measure of a global evaluation, older participants, and multiple informants ($0.24 + 0.10 + 11 + 0.08 + 0.07 = 0.60$). These predicted values are slightly higher than the two actually observed self-informant correlations that fulfill these criteria, $r = 0.49$ and 0.54 . Thus, the moderator effects should be seen as rough estimates of the effect of these variables on self-informant agreement. The data are insufficient to provide more precise estimates or to test more complex interactions among moderators.

Table 1 Studies and self-informant correlations included in the meta-analysis

Authors	N	Sample #	S-I correlation	Construct	Age group	Informant type	# Of informants	# Items (S-R)	# Items (I-R)
Barsky et al. (2004)	288	1	0.42	LS	O	FR/FA	1	M	M
	288	1	0.34	LS	O	FR/FA	1	S	M
	288	1	0.44	LS	O	FR/FA	1	S	M
Bassett et al. (1990)	639	2	0.50	LS	O	FR/FA	1	M	M
	538	3	0.48	PA	O	FR/FA	1	M	M
	122	4	0.31	PA	Y	FR/FA	1	M	M
Borkenau and Mauer (2007)	122	4	0.29	NA	Y	FR/FA	1	M	M
	122	4	0.59	LS	Y	FR/FA	1	M	M
	122	4	0.42	HAP	Y	FR/FA	1	S	S
Dew and Huebner (1994)	222	5	0.48	LS	Y	FA	1	M	S
Diener et al. (1995)	212	6	0.53	PA	Y	FR/FA	4.1	M	M
	212	6	0.39	NA	Y	FR/FA	4.1	M	M
	160	7	0.26	NA	Y	FR	1	M	M
Eisenberg et al. (1994)	266	8	0.34	LS	Y	FA	1	M	M
	99	9	0.54	LS	Y	FA	1	M	S
	195	10	0.34	HAP	Y	FR	4	S	S
Gilligan and Huebner (2002)	82	11	0.40	LS	O	FA	1	M	M
	80	12	0.50	LS	Y	FA	1	M	S
	217	13	0.45	LS	O	FR/FA	1	M	M
Gillman and Huebner (1997)	217	13	0.39	HAP	O	FR/FA	1	S	S
	60	14	0.36	HAP	Y	FR	2	S	S
	117	15	0.64	HAP	O	FR/FA	1	M	S
Hartmann (1934)	120	16	0.45	HAP	O	FR/FA	1	M	S
	95	17	0.63	HAP	O	FR/FA	1	M	S
	82	11	0.40	LS	O	FA	1	M	M
Heller et al. (2006)	80	12	0.50	LS	Y	FA	1	M	S
Huebner et al. (2002)	217	13	0.45	LS	O	FR/FA	1	M	M
	217	13	0.39	HAP	O	FR/FA	1	S	S
	60	14	0.36	HAP	Y	FR	2	S	S
Judge and Locke (1993)	117	15	0.64	HAP	O	FR/FA	1	M	S
	120	16	0.45	HAP	O	FR/FA	1	M	S
	95	17	0.63	HAP	O	FR/FA	1	M	S
Kammann et al. (1984)	60	14	0.36	HAP	Y	FR	2	S	S
Kozma and Stones (1988)	117	15	0.64	HAP	O	FR/FA	1	M	S
	120	16	0.45	HAP	O	FR/FA	1	M	S
	95	17	0.63	HAP	O	FR/FA	1	M	S

Table 1 continued

Authors	N	Sample #	S-I correlation	Construct	Age group	Informant type	# Of informants	# Items (S-R)	# Items (I-R)	
Lepper (1998)	971	18	0.45	PA	O	FR/FA	1	M	M	
	971	18	0.43	NA	O	FR/FA	1	M	M	
	971	18	0.56	LS	O	FR/FA	1	M	M	
	971	18	0.59	HAP	O	FR/FA	1	M	M	
	528	19	0.43	PA	O	FR/FA	1	M	M	
	528	19	0.33	NA	O	FR/FA	1	M	M	
	528	19	0.51	LS	O	FR/FA	1	M	M	
	528	19	0.54	HAP	O	FR/FA	1	M	M	
	212	20	0.43	PA	Y	FR/FA	3	M	M	
Lucas et al. (1996)	212	20	0.26	NA	Y	FR/FA	3	M	M	
	212	20	0.48	LS	Y	FR/FA	3	M	M	
	109	21	0.41	PA	Y	FR/FA	2	M	M	
	109	21	0.44	NA	Y	FR/FA	2	M	M	
	109	21	0.52	LS	Y	FR/FA	2	M	M	
	59	22	0.66	HAP	Y	FR	1	M	M	
	68	23	0.65	HAP	Y	FR	1	M	M	
	43	24	0.41	HAP	Y	FR	1	M	M	
	528	25	0.44	HAP	O	FA	1	M	M	
Lyubomirsky and Lepper (1999)	60	26	0.49	LS	Y	FR	2	M	M	
	70	27	0.42	LS	Y	FR	2	M	M	
	72	28	0.27	LS	Y	FR	2	S	M	
	72	28	0.43	HAP	Y	FR	2	S	S	
	73	29	0.28	LS	Y	FR	2	S	M	
	73	29	0.25	HAP	Y	FR	2	S	S	
	76	30	0.35	LS	Y	FR	2	S	M	
	Pavot and Diener (1993b)									

Table 1 continued

Authors	N	Sample #	S-I correlation	Construct	Age group	Informant type	# Of informants	# Items (S-R)	# Items (I-R)
Pavot et al. (1991)	76	30	0.33	HAP	Y	FR	2	S	S
	78	31	0.26	LS	Y	FR	2	S	M
	78	31	0.27	HAP	Y	FR	2	S	S
	89	32	0.65	PA	Y	FR/FA	7	S	S
	89	32	0.43	NA	Y	FR/FA	7	S	S
	89	32	0.64	LS	Y	FR/FA	7	M	M
	38	33	0.54	LS	O	FR/FA	3	M	M
	38	33	0.49	LS	O	FR/FA	3	M	M
	38	33	0.56	HAP	O	FR/FA	3	S	S
	104	34	0.43	PA	Y	FA	1	M	M
Phillips et al. (2002)	104	34	0.06	NA	Y	FA	1	M	M
	291	35	0.37	NA	O	FA	1	M	M
Pruchno et al. (2006)	126	36	0.48	HAP	Y	FA	3	S	S
Sandvik et al. (1993)	126	36	0.50	HAP	Y	FR	3	S	S
Schimmack and Diener (2003)	141	37	0.38	PA	Y	FR/FA	4	M	M
	141	37	0.39	NA	Y	FR/FA	4	M	M
	141	37	0.55	LS	Y	FR/FA	4	M	M
	70	38	0.45	LS	Y	FR/FA	3	M	M
Steger et al. (2006)	150	39	0.29	PA	Y	FR	1	M	M
	150	39	0.46	NA	Y	FR	1	M	M
	150	39	0.34	LS	Y	FR	1	M	M
Walker and Schimmack (2008)	150	40	0.29	PA	Y	FR	4	M	M
	558	41	0.30	PA	Y	FR	1	M	M
	558	41	0.20	NA	Y	FR	1	M	M
	272	42	0.33	PA	Y	FR	1	M	M
Watson and Clark (1991)	272	42	0.22	NA	Y	FR	1	M	M
	272	42	0.22	NA	Y	FR	1	M	M

Table 1 continued

Authors	<i>N</i>	Sample #	S-I correlation	Construct	Age group	Informant type	# Of informants	# Items (S-R)	# Items (I-R)
	148	43	0.34	PA	O	FA	1	M	M
	148	43	0.29	NA	O	FA	1	M	M
Watson and Humrichouse (2006)	301	44	0.38	PA	O	FA	1	M	M
	301	44	0.39	NA	O	FA	1	M	M
	301	43	0.27	PA	O	FA	1	M	M
	301	43	0.32	NA	O	FA	1	M	M

S-I Correlations are Pearson correlations of self and informant ratings. Construct was coded as follows: PA positive affect, NA negative affect, LS life satisfaction, HAP happiness; O older target participant (>24 years of age), Y younger target participant (<24 years of age); FR/FA friend and/or family, FR/FA friend and/or family, FA family; M multiple-item scale, S single-item scale

Table 2 Regression analysis

Parameter	Estimate	Lower bound	Upper bound
Intercept/Positive affect	0.24	0.13	0.35
Evaluation	0.11	0.04	0.19
Negative affect	-0.06	-0.11	-0.02
Life satisfaction	-0.04	-0.10	0.02
Age	0.08	0.02	0.15
# Informants	0.07	0.00	0.13
Type informant	0.02	-0.07	0.11
# Items	0.10	0.00	0.20
Publication year	-0.01	-0.04	0.03

Note: 95% confidence interval. Predictor variables: evaluation (0 = positive affect/negative affect vs. 1 = life-satisfaction, happiness), negative affect (0 = positive affect, life-satisfaction, happiness, 1 = negative affect), life satisfaction (0 = positive affect, negative affect, happiness 1 = life satisfaction), age [0 = young (<24), 1 = older (>24)], # informants (0 = single, 1 = multiple), type informant (0 = friend, 0.5 = friend/family, 1 = family), # items (0 = single, 0.5 mixed, 1 = multiple), publication year (decades as units)

6 Discussion

A science of well-being requires valid measures of well-being. Although a common definition of well-being is lacking, global ratings of happiness or life-satisfaction, and the balance of positive affect versus negative affect are commonly used indicators of well-being. For these indicators it is possible to estimate their validity as measures of life-satisfaction, happiness, positive affect, or negative affect by examining convergent validity across independent methods. The most widely used methods are self-ratings and informant ratings. Our meta-analysis revealed an average convergent validity of $r = 0.42$, with a relatively tight credibility interval ranging from 0.39 to 0.45. This finding confirms the common assumption in the literature that self-ratings of well-being have some validity. Moreover, we found some significant moderators of self-informant agreement. Agreement increased with age, number of informants, and number of items. Most importantly, self-informant agreement varied as a function of the construct being assessed. Affect measures revealed lower self-informant agreement than global judgments and negative affect produced less agreement than positive affect. This finding suggests that studies that rely on global judgments of life satisfaction and happiness as indicators of well-being may produce more valid results than studies that rely on affective indicators. However, firm conclusions are not possible because our meta-analysis does not provide independent estimates of the validity of self-ratings and informant ratings. It is possible that lower agreement for affective measures is due to lower validity in informant ratings of affect. In this case, self-ratings of affect might be as valid as self-ratings of life-satisfaction and global happiness.

Our meta-analysis has a number of limitations. First, the evidence base is rather limited, especially to obtain precise estimates of moderator effects. Second, self-informant correlations do not reveal the amount of valid variance in self-ratings and informant ratings. For this purpose, more than two methods are needed. Multiple informants, especially those without direct contact to each other, could be used to increase the number of methods. Unfortunately, most studies with multiple informants simply averaged across informants. Future studies should treat each informant as a separate variable and examine the correlations among all raters with a latent variable model. Finally, our conclusions regarding

moderator variables have to be treated with caution because effects can be contaminated by unobserved factors. More carefully designed studies that hold other factors constant are needed to examine whether the observed effects are reliable. Moreover, some moderators may have effects even if they were not significant in our analysis due to low power. Finally, our estimates of validity are limited to the validity of indicators as measures of the intended construct (e.g., self-ratings of life-satisfaction as measure of life-satisfaction). How valid the measures are as measures of well-being depends on the definition of well-being and the unknown empirical relation between well-being and the constructs examined in the meta-analysis (e.g., life-satisfaction). As this relation can at best be 1.00, the validity estimates in this study are estimates of the upper limit for the validity as measures of well-being. Overall, the validity of these measures is moderate and there is no evidence to suggest that a single indicator is considerably more valid than other indicators. Our results also provide no empirical support for the assumption that self-ratings on global evaluative measures can be treated as unbiased measures of wellbeing, quality of life, or utility. One solution to increase the modest validity of well-being measures is to use multiple indicators to measure well-being. For example, a measure based on reports of multiple indicators by multiple informants is likely to be more valid than a measure based on a single indicator by a single rater (Walker and Schimmack 2008).

Another noteworthy finding was that self-informant agreement has not significantly increased over time since Hartmann (1934) conducted the first test of convergent validity. This finding shows that well-being science has not made any progress in creating more valid measures of well-being. The reason for this lack of progress is the common tendency to equate the construct with a measure and take observed findings at face value (Borsboom 2006). More precise measurement has been a driving force of scientific discovery in other sciences, like astronomy. It is time to take measurement more seriously and to devote more attention to the creation of better well-being measures.

Acknowledgments We thank Simone Walker, Naoki Nakazato, and the members of the well-being laboratory at the University of Toronto Mississauga for their valuable comments.

References

References marked with an asterisk indicate studies included in the meta-analysis

- *Barsky, A., Thoresen, C. J., Warren, C. R., & Kaplan, S. A. (2004). Modelling negative affectivity and job stress: A contingency-based approach. *Journal of Organizational Behavior*, 25, 915–936.
- *Bassett, S. S., Magaziner, J., & Hebel, J. R. (1990). Reliability of proxy response on mental health indices for aged, community-dwelling women. *Psychology and Aging*, 5, 127–132.
- *Borkenau, P., & Mauer, N. (2007). Well-being and the accessibility of pleasant and unpleasant concepts. *European Journal of Personality*, 21, 169–189.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425–440.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cohen, J. (1994). The earth is round (P -less-than .05). *American Psychologist*, 49, 997–1003.
- Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment*, 15, 110–117.
- *Dew, T., & Huebner, E. S. (1994). Adolescents' perceived quality of life—an exploratory investigation. *Journal of School Psychology*, 32, 185–199.

- Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, 95, 542–575.
- *Diener, E., Smith, H., & Fujita, F. (1995). The personality structure of affect. *Journal of Personality and Social Psychology*, 69, 130–141.
- Eid, M., & Diener, E. (2004). Global judgments of subjective well-being: Situational variability and long-term stability. *Social Indicators Research*, 65, 245–277.
- *Eisenberg, N., et al. (1994). The relations of emotionality and regulation to dispositional and situational empathy-related responding. *Journal of Personality and Social Psychology*, 66, 776–797.
- *Gilligan, T. D., & Huebner, E. S. (2002). Multidimensional life satisfaction reports of adolescents: A multitrait-multimethod study. *Personality and Individual Differences*, 32, 1149–1155.
- *Gilman, R., & Huebner, E. S. (1997). Children's reports of their life satisfaction—convergence across raters, time and response formats. *School Psychology International*, 18, 229–243.
- *Hartmann, G. W. (1934). Personality traits associated with variations in happiness. *Journal of Abnormal and Social Psychology*, 29, 202–212.
- *Heller, D., Watson, D., & Ilies, R. (2006). The dynamic process of life satisfaction. *Journal of Personality*, 74, 1421–1450.
- *Huebner, E. S., Brantley, A., Nagle, R. J., & Valois, R. F. (2002). Correspondence between parent and adolescent ratings of life satisfaction for adolescents with and without mental disabilities. *Journal of Psychoeducational Assessment*, 20, 20–29.
- *Judge, T. A., & Locke, E. A. (1993). Effect of dysfunctional thought processes on subjective well-being and job satisfaction. *Journal of Applied Psychology*, 78, 475–490.
- *Kammann, R., Smith, R., Martin, C., & McQueen, M. (1984). Low accuracy in judgments of others' psychological well-being as seen from a phenomenological perspective. *Journal of Personality*, 52, 107–123.
- *Kozma, A., & Stones, M. J. (1988). Social desirability in measures of subjective well-being—age comparisons. *Social Indicators Research*, 20, 1–14.
- *Lepper, H. S. (1998). Use of other-reports to validate subjective well-being measures. *Social Indicators Research*, 44, 367–379.
- *Lucas, R. E., Diener, E., & Suh, E. (1996). Discriminant validity of well-being measures. *Journal of Personality and Social Psychology*, 71, 616–628.
- *Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, 46, 137–155.
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.
- Pavot, W., & Diener, E. (1993). Review of the satisfaction with life scale. *Psychological Assessment*, 5, 164–172.
- *Pavot, W., & Diener, E. (1993). The affective and cognitive context of self-reported measures of subjective well-being. *Social Indicators Research*, 28, 1–20.
- *Pavot, W., Diener, E., Colvin, C. R., & Sandvik, E. (1991). Further validation of the satisfaction with life scale—evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment*, 57, 149–161.
- *Phillips, B. M., Lonigan, C. J., Driscoll, K., & Hooe, E. S. (2002). Positive and negative affectivity in children: A multitrait-multimethod investigation. *Journal of Clinical Child and Adolescent Psychology*, 31, 465–479.
- *Pruchno, R. A., Lemay, E. R., Field, L., & Levinsky, N. G. (2006). Predictors of patient treatment preferences and spouse substituted judgments: The case of dialysis continuation. *Medical Decision Making*, 26, 112–121.
- *Sandvik, E., Diener, E., & Seidlitz, L. (1993). Subjective well-being: The convergence and stability of self-report and non-self-report measures. *Journal of Personality*, 61, 317–342.
- Schimmack, U. (2008). Wellbeingscience.org—the science of wellbeing (2008, September 25). Retrieved October 20, 2008, from http://www.erin.utoronto.ca/~w3psyuli/WellBeingScience/wellbeing_science.htm.
- *Schimmack, U., & Diener, E. (2003). Predictive validity of explicit and implicit self-esteem for subjective well-being. *Journal of Research in Personality*, 37, 100–106.
- Schimmack, U., & Oishi, S. (2005). The influence of chronically and temporarily accessible information on life satisfaction judgments. *Journal of Personality and Social Psychology*, 89, 395–406.
- Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 61–84). New York: Russell Sage Foundation.
- *Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology*, 53, 80–93.

- *Walker, S. S., & Schimmack, U. (2008). Validity of a happiness implicit association test as a measure of subjective well-being. *Journal of Research in Personality, 42*, 490–497.
- *Watson, D., & Clark, L. A. (1991). Self-versus peer ratings of specific emotional traits: Evidence of convergent and discriminant validity. *Journal of Personality and Social Psychology, 60*, 927–940.
- *Watson, D., Hubbard, B., & Wiese, D. (2000). Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology, 78*, 546–558.
- *Watson, D., & Humrichouse, J. (2006). Personality development in emerging adulthood: Integrating evidence from self-ratings and spouse ratings. *Journal of Personality and Social Psychology, 91*, 959–974.
- Whitener, E. M. (1990). Confusion of confidence-intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75*(3), 315–321.
- Wilson, W. (1967). Correlates of avowed happiness. *Psychological Bulletin, 67*, 294–406.