Available online at www.sciencedirect.com
ScienceDirect
Social Science Research 35 (2006) 1120–1131
Social Science RESEARCH
www.elsevier.com/locate/ssresearch

Reply

Bad news indeed for Ryff's six-factor model of well-being ☆

Kristen W. Springer [a,b,*], Robert M. Hauser [a], Jeremy Freese [a,c]

[a] *Department of Sociology and the Center for Demography of Health and Aging, University of Wisconsin-Madison, USA*
[b] *Department of Sociology, Rutgers University, USA*
[c] *Robert Wood Johnson Foundation Scholars in Health Policy Research, Harvard University, USA*

Available online 20 March 2006

Abstract

Springer and Hauser (An Assessment of the Construct Validity of Ryff's Scales of Psychological Well-Being: Method, Mode, and Measurement Effects. 2006. Social Science Research 35) tested one key aspect of the validity of Ryff's six-factor model of psychological well-being (RPWB), namely, whether there is substantial independent variation among the six factors. In several large and heterogeneous samples, under a variety of model specifications, and using various sets of RPWB items, we found very high factor correlations among the dimensions of well-being, especially personal growth, purpose in life, self-acceptance, and environmental mastery. That is, the six-factor model makes theoretical claims that do not yield large or consistent empirical distinctions when standard measures and instrumentation are used. Where Ryff and Singer's comment (Best News Yet on the Six-Factor Model of Well-Being. 2006. Social Science Research 35) refers directly to that analysis, their

DOI of original article: 10.1016/j.ssresearch.2006.01.002.

☆ The research reported herein was supported by the National Institute on Aging (R01 AG-9775 and P01 AG-21079), by the William Vilas Estate Trust, by the Robert Wood Johnson Foundation, and by the Graduate School of the University of Wisconsin-Madison. Computation was carried out using facilities of the Center for Demography and Ecology at the University of Wisconsin-Madison, which are supported by Center and Training Grants from the National Institute of Child Health and Human Development and the National Institute on Aging. We thank Sheung-Tak Cheng for sharing the factor correlations from his 2005 *Personality and Individual Differences* paper. We also thank Richard T. Campbell, Seth M. Hauser, Taissa S. Hauser, Tetyana Pudrovska, James Raymo, Halliman H. Winsborough, James Yonker, and Zhen Zeng for helpful advice. The opinions expressed herein are those of the authors.

* Corresponding author. Department of Sociology, Rutgers University, 54 Joyce Kilmer Avenue, Piscataway, NJ 08854, USA (After July 2006). Fax: +1 608 265 5389.
E-mail address: kspringe@ssc.wisc.edu (K.W. Springer).

0049-089X/$ - see front matter
doi:10.1016/j.ssresearch.2006.01.003

methodological discussion is most often irrelevant or incorrect. Their text largely ignores and fails to challenge our strong empirical findings about the factorial structure of well-being. In this response, we reinforce these findings and their implications for the (in)validity of the six-factor well-being model as implemented by Ryff. We also explain why Ryff and Singer's lengthy review of studies that show differential relationships of RPWB factors with other variables should be interpreted with far greater caution than Ryff and Singer recognize. We offer recommendations for analyzing RPWB items in surveys that have already been conducted, but we also emphasize the need for a thorough rethinking of the measurement and dimensionality of psychological well-being.

Keywords: Psychological well-being; Well-being; Measurement; Survey design; Confirmatory factor model; Statistical power

1. Introduction

The main finding of Springer and Hauser (2006) is that, in self-administered survey instruments, estimated correlations among four of the six latent dimensions of Ryff's model of psychological well-being are so close to 1.0 that there are scant meaningful empirical differences among them. The finding holds in two national samples (MIDUS and NSFH II) and in two large regional samples (WLS graduates and siblings).[1] The finding holds despite differences in the items used in the different samples. In the WLS, the correlations among the four factors are all above 0.9 even before we adjust for method artifacts created by item proximity and polarity (reverse-scoring).

Our findings demonstrate the need for constructive and thorough reconsideration of Ryff's measurement scales and of the six-factor model of psychological well-being.[2] They imply either that the measures used are inadequate to capture the distinctions intended by the theoretical model, that the distinctions themselves are incoherent or inconsequential, or both—at least as applied to general population samples. While our analyses cannot adjudicate among these possibilities, our results should be regarded as unambiguous bad news for the continued broad use of these indistinct subscales in the study of psychological well-being.

By contrast, Ryff and Singer (2006) claim that our findings are the "best news yet" for the six-factor model of psychological well-being and its current methods of instrumentation. This is wishful thinking. As the subtleties of factor analytic results may be easily misunderstood or viewed as a technical exercise without practical import, let us illustrate one upshot of our findings by imagining two researchers who are working with the 28 RPWB items from the 1993 WLS graduate data that were intended to measure personal growth, purpose in life, self-acceptance, and environmental mastery. Researcher A dutifully constructs four seven-item scales based on Ryff's theoretical scheme. Researcher B *pays absolutely no attention to what the items were intended to measure*, but instead constructs four new seven-item scales based only on the matters Ryff and Singer suggest were "muddled" by us: Whether the items appear earlier or later in the instrument and whether they are positively or negatively scored. That is, Researcher B's four scales are (1) positively scored

[1] Hauser et al. (2005) have cross-validated the findings in longitudinal analyses of NSFH and WLS data.

[2] For convenience, we use "we" and "our" throughout this reply in reference to Springer and Hauser (2006), even though Jeremy Freese was not an author of that paper.

and early in the instrument, (2) positively scored and later in the instrument, (3) reverse-scored and early in the instrument, and (4) reverse-scored and later in the instrument. The reliabilities (Cronbach's α) of the four scales produced by these two analysts are virtually identical (0.767 for A vs. 0.771 for B). The average correlations among the scales are also virtually identical (0.611 for A vs. 0.616 for B). In other words, *the items comprising the four theoretically defined subscales are so interchangeable in these data that one does as well with scales constructed by rearranging items according to non-substantive criteria that are not supposed to matter*.[3] This is bad news indeed for Ryff's six-factor model of well-being.

Ryff and Singer's reinterpretation of our findings as the "best news yet" for current theory and practice is based on three faulty counterarguments. First, they emphasize that we found a six-factor solution best fit our data using the criterion of nominal statistical significance. In no way does this finding over-ride the strong implications of the very high factor correlations among dimensions. In addition, this argument carries little weight because models based on non-substantively directed rearrangements of items—like those by Researcher B above—yield superior fit. Second, they regard several of our methodological decisions as misguided. These criticisms are easily countered, and the warrant for our adjusted estimates of correlations among factors is readily demonstrated. Third, Ryff and Singer devote more than half of their comment to a review of studies that purportedly find different dimensions of well-being have differential relationships with other variables. We regard this as a particularly dangerous evidentiary foundation on which to rest confidence in the dimensional structure of a construct. Their review is largely consistent with what one would expect to find even if several measures of putatively separate dimensions of well-being were, in fact, just different measures of exactly the same thing. We discuss each of these points in turn.

2. Model estimates matter

Ryff and Singer declare that the "key take-home message" of our study is that a six-factor solution was found to fit RPWB items better than more parsimonious alternatives. Not only do they assert that this is the only consequential result of the study, but they suggest that the rest of the paper is little more than a "lengthy exercise" directed toward "trying to discredit what their own analyses show" (p. 1103). This is an egregious misreading of our text. Our key take-home message relies on the simple idea that estimates from statistical models matter. The message is that the estimated correlations among four of the factors are well in excess of 0.9. These high correlations demonstrate the need to rethink the substantive distinctions intended by those four factors and/or their measurement. In short, a reader who takes home only that a six-factor model fits best has missed the most important empirical findings and substantive implications of our paper.[4]

[3] In the WLS, the four PWB scales for these factors include two more negatively phrased (reverse-scored) items than positively phrased items. Thus, to construct four seven-item scales based on proximity and polarity, it was necessary to combine two negatively phrased items with five positively phrased items in one scale. Alternatively, one can construct two six-item scales of only positively phrased items and two eight-item scales of negatively phrased items. The results are virtually identical (average $\alpha = 0.768$, average intercorrelation of scales $= 0.604$).

[4] In this reply, we focus on well-being as measured in self-administered instruments. Springer and Hauser (2006) contained extensive comparisons of self-administered vs. telephone or other personal administrations of RPWB items, and they discussed evidence that self-administered batteries have greater validity. As Ryff and Singer's (2006) comment reply makes no mention of this mode effect, we presume they have little quarrel with it.

Although we wish to emphasize that the important issue here is not model fit but the very high correlations among four of the RPWB factors, readers should also keep in mind the severe limitations of Ryff and Singer's claim that a six-factor model is needed to fit the data that we analyzed. As we noted, when one has a very large sample, like the WLS graduates ($N = 6282$) or NSFH II ($N = 9240$), one should expect to reject a model with fewer factors, even when the correlations between factors in that model are close to one. Given a large enough sample, no restriction on parameters is likely to fit using criteria of nominal statistical significance or even a penalized test statistic (Raftery, 1995; Weakliem, 1999). In the WLS graduate sample, a model that treats personal growth, purpose in life, self-acceptance, and environmental mastery as forming a single factor fits *very* well using another widely accepted measures of fit, the root mean square error of approximation (RMSEA = 0.035) (Loehlin, 2004, pp. 67–70). Similarly, when we estimate this same model with the NSFH II data, RMSEA = 0.047.[5] Furthermore, even with a moderately large sample, like MIDUS ($N = 2731$ in our analysis), the six-factor model can be rejected in favor of a simpler alternative: A model that does not distinguish between purpose in life and personal growth fits the MIDUS data better than the six-factor model; that is, it has a lower BIC statistic (−257 vs. −230). In other words, although Ryff and Singer wish to add our study to the list of confirmations of the six-factor model, it is not true that six factors are required to fit the data that we analyzed. When fit is assessed by criteria other than simple statistical significance, we find that more parsimonious models fit well and sometimes better.

Moreover, Ryff and Singer offer an unconvincing discussion of other studies of factorial structure in support of the six-factor model of well-being. Specifically, Clarke et al. (2001) and Ryff and Keyes (1995) use telephone or in-home administration of RPWB items—a technique demonstrated to be less valid for RPWB measurement than self-administration. The other cited studies use self-administered instruments, and find scant evidence supporting a six-factor model of well-being. Kafka and Kozma (2002), for example, conclude their paper saying "it would appear that the structure of [RPWB] is limited to face validity. (p. 186)" Van Dierendonck (2004) also uses self-administered RPWB items and finds factor correlations approaching 0.90 among self-acceptance, environmental mastery, purpose in life and personal growth. Cheng and Chan (2005) analyze a Chinese version of RPWB using product-moment covariances with maximum likelihood estimation and find factor correlations between 0.69 and 0.93 among these same four factors.[6] These studies provide no evidence favoring the six-factor model of well-being.

Again, the main issue is the very high correlations among RPWB factors, not model fit. Ryff and Singer interpret at least part of our concern about these high correlations as reflecting an ignorance of the logic by which the scales were developed—the construct-oriented approach to scale development. Given our strong empirical evidence that four of the

[5] RMSEA < 0.05 is regarded as providing very good fit to a covariance structure (Loehlin, 2004, pp. 68–73).

[6] Neither Van Dierendonck (2004) nor Cheng and Chan (2005) report these correlations. Each paper reports on confirmatory factor models estimated from a (product-moment) covariance matrix, which accounts, along with the use of maximum likelihood estimation, for the lower range of estimated factor correlations. Analyzing polychoric correlations with weighted least squares estimation is the preferred strategy for ordinal data, like RPWB (Jöreskog and Dag Sörbom, 1996a,b). As demonstrated in our paper, this strategy yields higher factor correlations. Van Dierendonck sent his factor correlations to us on 7/26/05, and Cheng sent them to us on 1/05/06.

subscales are not empirically distinct, it does not much matter how they were developed. However, we very well understand the construct-oriented approach to scale development. Central to it, of course, is retaining items that correlate highly with other items that purport to measure the same thing and that do not correlate highly with items that measure different things. That is, the procedure should yield both convergent and discriminant validity. We were concerned that excessive reliance on correlational evidence might produce the appearance of convergent or discriminant validity for the wrong reason, that is, because of singular characteristics of items. That concern was initially prompted by Ryff and Keyes (1995, p. 720) choice of three indicators from among 20 indicators of each subscale that had already been screened for convergent and divergent validity. Evidently, these worries were groundless, as demonstrated by the very high correlations among factors estimated from the 18 indicators in the MIDUS and NSFH II data. Indeed, this leads to another "take home message": Given Ryff and Singer's description of admirably extensive and iterative efforts to measure six distinct dimensions of well-being, it is remarkable that four of the six measures are virtually indistinguishable in several independent samples.

3. Why our models are right

Ryff and Singer object to the methodological corrections we make for adjacent items, items that are "reverse-scored," and seemingly redundant items. Their point regarding redundant items is arguable, but it ends up of little consequence for our estimated correlations among RPWB factors. Regarding their objections on the other two points, we should first make plain that correcting for method artifacts is not itself a criticism of the instrumentation, as Ryff and Singer seem to believe. One should adjust for method artifacts when one has good reason to think that they exist. Estimates are improved by taking them into account. It is wrong to think that such adjustments should be reserved for instances of mistaken practice in instrumentation. We might better have used "adjusted" rather than "corrected" to allay this confusion. In any event, we agree with Ryff and Singer's observations about the benefits of mixed item-ordering and reverse-scored items, but they are wholly beside the point here. The important thing is that each of the adjustments—for adjacent and for reverse-scored items—substantially and consequentially improves the fit of our models in the WLS, MIDUS, and NSFH II, vindicating our supposition that such adjustments are warranted while undermining Ryff and Singer's claim that the adjusted findings should be ignored.

Take first the adjustment for similar responses to adjacent items. Ryff and Singer dispute the idea that survey participants may be influenced more by their response to the immediately preceding item than by responses to other items. The empirics here are plain: In the WLS mail survey, among measures of theoretically distinct constructs, the average correlation between adjacent items is $|r| = 0.253$, while the average correlation between nonadjacent items is $|r| = 0.224$. In most cases, disturbances are positively correlated when adjacent items have the same polarity (i.e., neither or both are reverse-scored), and negatively correlated when adjacent items have different polarity. Participants tended disproportionately to choose the same response category from one item to the next. While correlating the errors of adjacent items has modest effects on estimated correlations among the RPWB factors, it substantially improves model fit. Again, Ryff and Singer ignore the strong empirical evidence that these correlations exist in each of the samples that we analyzed.

Ryff and Singer also object to the adjustment for reverse-scored items. They are right that we might better have chosen the term "reverse-scored" for "negatively worded" in describing 22 items in the 1993 WLS mail survey and 8 items in the MIDUS and NSFH II surveys that were permitted to load on a correlated methods factor in some models.[7] To avoid confusion, we followed the terminology that Ryff and Keyes (1995, p. 720) used to describe reverse-scoring in the much-used set of 18 Ryff scale items: "Each scale included both positively and negatively phrased items." In any event, while we agree that it is wise to include reverse-scored items in scales in order to avoid acquiescence bias, we are not alone in thinking it highly plausible that reversals in coding can produce a correlated methods factor. As DeVellis (2003, p. 69) writes in his book on scale development, "Reversals in item polarity may be confusing to respondents, especially when they are completing a long questionnaire."[8] Once more, the empirics are clear. Introduction of a factor for negatively phrased items improves fit substantially in each set of data. In the WLS, for items intended as measures of different constructs, the average correlation among items with the same item polarity is $|r| = 0.263$, while the correlation among items with different item polarity is $|r| = 0.190$.

In sum, the adjustments for item proximity and polarity are substantively justified, and they are plainly supported by our empirical analysis. We find it ironic that Ryff and Singer focus on model fit in defending the six-factor model against simpler alternatives, while dismissing clear evidence that adjustments for method effects improve fit.

Because Ryff and Singer recommend so strongly that researchers focus only on which model fits the data best and dispute our adjustments for proximity and polarity effects in estimating factor correlations, we elaborate our earlier example of measures based only on those two characteristics of the items. For the measures of personal growth, purpose in life, self-acceptance, and environmental mastery, the best-fitting four-factor model *ignores* the intended theoretical distinctions among items. Rather, models with four subscales based solely on item proximity and polarity fit *better* than a model based on Ryff's theoretically determined subscales. This is true for all three datasets that we analyzed: WLS ($L^2 = 4568.6$ vs. 4894.6 with 344 df), MIDUS ($L^2 = 861.5$ vs. 912.0 with 48 df), and NSFH II ($L^2 = 2308.3$ vs. 2639.5 with 48 df).[9]

Once again, we disagree with the proposition that, in evaluating an *n*-factor model, the "take-home message" should be defined solely by whether an *n*-factor model fits best. However, if one does wish to focus on this narrow criterion, one might at least expect that the best-fitting *n*-factor model should be one that actually represents the distinctions intended by the theory, not one that ignores them completely.

[7] Referring to Table 2b in Springer and Hauser (2006), these are WLS mail items 3, 4, 6, 8, 9, 12, 15, 16, 17, 18, 23, 24, 25, 26, 27, 31, 33, 35, 36, 38, 40, and 42; MIDUS items 4, 5, 6, 7, 10, 14, 15, and 16; and NSFH II items 3, 4, 5, 7, 8, 10, 12, and 14.

[8] Also, "Positively and Negatively Worded Items" is the heading of the relevant section of DeVellis's (2003) text on scale development.

[9] Recall that the order of presentation of the 18 items differs between the MIDUS and NSFH II data. Also, for these data, we report test statistics in the text only for the *worst*-fitting of four methodologically specified models, which differ in the assignment of items to positively scored and reverse-scored factors. In MIDUS, the best-fitting model in this set yields the likelihood-ratio chi-square statistic of 692.7, and in NSFH II, the best-fitting model yields $L^2 = 2204.9$. The WLS has even numbers of positively scored and reverse-scored items, and, thus, we did not need to test a variety of methodologically specified four-factor models for the WLS. Details are available from the authors.

4. The review of differential correlation is deeply flawed

What is the practical import for researchers that the standard instrumentation of RPWB yields such high estimated correlations among four of the six factors that there is little substantive difference among them? A non-obvious consequence is that researchers will still find relationships with other variables that are significant for one or two of these four subscales, but not others. It will then be easy for researchers to spin retrodictive stories about why this should be so, when the different patterns of significance mostly reflect the mundane statistical consequence of having multiple measures of essentially the same construct. The problem will be exacerbated further when researchers create indexes based on smaller numbers of items, say three or four per scale. The correlations among subscales will not be exceptionally large because the indexes all have moderate reliabilities at best. In other words, *researchers will wrongly believe they are generating knowledge about specific components of well-being*—and even that they are validating the six-factor model—when actual knowledge would instead be both strengthened and simplified by an improved understanding of the dimensionality of well-being.

Given this concern, it is interesting that at least half of Ryff and Singer's comment on our paper—four of the five "types of evidence" that they present—does not offer direct evidence about factorial structure but instead offers inferences from relationships between dimensions of well-being and other variables. The set of correlates that Ryff and Singer discuss is commendably broad—age, time, or other psychological variables; social and demographic variables; biological variables; and clinical interventions. The argument in each case is the same, that differential relationships between RPWB subscales and some other variables validate the claim that the subscales are empirically distinct.[10] No less than five times do Ryff and Singer offer italicized statements like, "*no two or three PWB scales showed the same pattern of effects*" (p. 1108), where the italics presumably convey their opinion that such findings represent especially persuasive evidence for their position.

Without exception, Ryff and Singer do not report that a relationship of one variable with an RPWB subscale is significantly different from that of another variable, but only that results are significant for some variables, and not significant or—in very few instances—significant and of opposite sign for other variables.[11] The conclusions they draw from their literature review are deeply flawed because Ryff and Singer ignore the substantial probability of finding seemingly different relationships between RPWB subscales and other variables, even when the RPWB subscales measure exactly the same thing. Ryff and Singer would be able to provide roughly the same literature review even if four of the six RPWB factors were identical.

To document this point, we carried out multiple simulations to estimate statistical power in analytic situations that closely resemble studies of the correlates of RPWB subscales. Even for measures of exactly the same latent construct, using measures of modest reliability (a), correlations with outcomes (r), sample sizes (N), and numbers of comparisons (k; outcomes, subgroups, or outcome/subgroup combinations), one can easily have a

[10] Regarding the evidence Ryff and Singer present from clinical interventions, their discussion concerns the question of whether it is clinically useful to connect self-reports of patients' positive experiences with subdimensions of RPWB. While we applaud the usefulness of the dimensions in clinical treatment, their checklist of concepts is irrelevant to the existence of differential correlation.

[11] See the discussion below of age differences in psychological well-being.

very high probability of observing different patterns of significant results. For example, if $a = .50$, $r = .10$, $N = 1250$, and $k = 6$, then power analyses conducted by simulation indicate that, with four measures of the same latent construct, at least 79% of studies would show the different patterns of effects embraced by Ryff and Singer as evidence of substantive distinction among subscales. We have chosen these parameters to resemble those for an analysis of four RPWB subscales and three outcomes in a survey like MIDUS, assuming that the findings are compared between women and men. Several other simulations, approximating the research designs of articles cited by Ryff and Singer, also yield high probabilities that no two or three RPWB subscales would show the same pattern of effects.[12] In other words, the findings emphasized by Ryff and Singer as supporting distinctions among RPWB subscales are very likely to occur by chance alone when multiple subscales actually measure the same thing.

As power analyses are virtually absent from primary studies or reviews in this literature, we think that researchers may dramatically overstate the chance that findings reflect real, substantive differences among different subscales instead of just the statistical consequences of imperfect measurement. Importantly, this point holds even if one believes that the four measures capture real distinctions, but they are just very highly correlated.

Two key points here deserve repeated emphasis. First, saying that, "X is significantly correlated with Y_1 and not with Y_2," does not tell us that there is any significant difference between the correlations of X with Y_1 and X with Y_2. Second, a significant difference between correlations does not in itself tell us that the several supposed subdimensions of psychological well-being represent more than a single latent factor. Even in a more sophisticated and error-adjusted analysis that models the factorial structure, a finding of differential correlation would require evidence against a model of proportional change in the correlation of RPWB subscales with other variables (Hauser and Goldberger, 1971). But that point gets way ahead of the state of the evidence assembled by Ryff and Singer, for none of their evidence pertains to error-corrected measures of psychological well-being.

To illustrate these ideas more formally, consider the path diagram in the upper half of Fig. 1, which shows the effects of three X_j on three Y_i by way of an intervening construct, η. For present purposes, suppose that the Y_i are three well-being factors and the X_j are possible causes of well-being (but our argument holds equally if we reverse the direction of causation). In this top figure, the Y_i are empirically distinct, but they also share a single common factor, η. The model says that

$$\eta = \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \zeta,$$

$$Y_1 = \lambda_1 \eta + \varepsilon_1,$$

$$Y_2 = \lambda_2 \eta + \varepsilon_2,$$

$$Y_3 = \lambda_3 \eta + \varepsilon_3,$$

where ζ and the ε_i and independent stochastic disturbances. In this model, the nine effects of the X_j on the Y_i are given by the $\gamma_i \lambda_j$. That is, the statistical relationships between X_j and Y_i are strictly proportional, but they need not be "identical" when there is only

[12] Stata 9 code and additional details of these simulations are available from the authors.

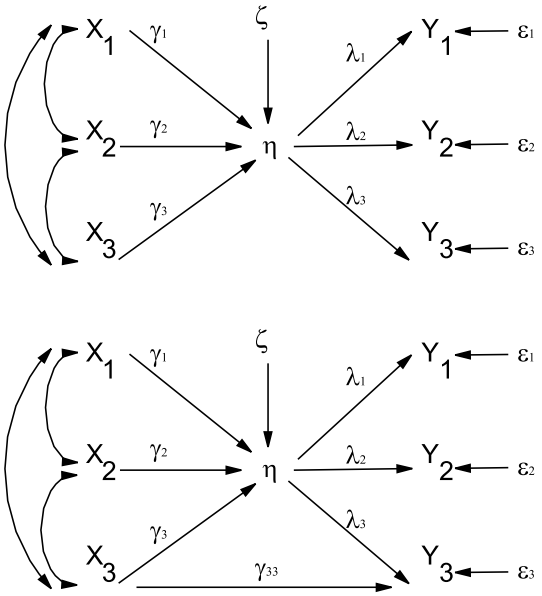

Fig. 1. Two multiple indicator, multiple cause (MIMIC) models of differential correlation.

one dimension to the relationships between psychological well-being and its antecedents. This model provides an appropriate null hypothesis for Ryff and Singer's claim about differential correlation. Correlations between the Y_i and X_j could differ because the λ_i differ or because the variances of the ε_i differ, even when there is only one dimension of psychological well-being. That is, differences in correlations between subdimensions of well-being and another variable need not imply differential relationships that would invalidate a single-factor model of psychological well-being.

The path diagram at the bottom of the figure displays a violation of the proportionality assumption; that is, because $\gamma_{33} \neq 0$, the relationships of X_1, X_2, or X_3 with Y_3 are not produced by a model of proportional effects. In this case, one would have to reject the idea that psychological well-being is one-dimensional. Note that, while changes of sign in estimated relationships between X_j and Y_i would appear to provide prima facie evidence that the model of proportional change has been violated, such findings may occur in a sample even when the model of proportional change holds in the population. Parallel ideas apply to the assessment of inter-population comparisons (or moderator effects) that are also discussed by Ryff and Singer, but we shall not elaborate them here.

In short, the appearance of differential correlation[13] among constructs with other outcomes should not be regarded as persuasive evidence about factorial structure without disciplined statistical analyses. Ryff and Singer merely recount a catalog of studies, none of which provides the requisite analysis. We invite readers to reconsider the literature review presented by Ryff and Singer and ask whether they seem to merely list one correlation after another, rather than offer a substantive, theoretically informed model that would discipline their use of evidence. In this respect, we are especially wary of

[13] We use "correlation" here in the generic sense, not with specific reference to correlation coefficients.

the biological evidence offered by Ryff and Singer, which is derived from very small samples.

Additionally, differential relationships should also be substantively important. For example, consistent with Ryff and Singer's review, we have also found statistically significant differences among the age trajectories of average subscale values. However, we find no strong age trajectories or large, consistent differences between subscales in change over time. Almost all of the variation in each subscale occurs within ages or periods. In analyses of change in RPWB among participants in the WLS from 1993 to 2004, we find that only 0.66 percent of the variance in personal growth occurs between years. In all of the five remaining dimensions, change accounts for less than 0.5 percent of the variance. Similarly, in the MIDUS survey, where Ryff and Singer make much of age variation, no more than two percent of the variance in any of the RPWB scales occurs across age groups (Pudrovska et al., 2005). We do not think much weight should be given to such weak evidence of differential correlation in assessing RPWB, even if it is a more disciplined type of evidence than the kind comprising the bulk of Ryff and Singer's literature review.

To be sure, there may be *real* instances of differential relationships between subdimensions of well-being and other variables. The problem with Ryff and Singer's review is that they fail to provide a disciplined treatment of the evidence. If the very high correlations that we find among subscales of RPWB are even approximately correct, many studies will still meet the weak standard of differential correlation guiding Ryff and Singer's literature review. At the very least, we urge caution in accepting specific findings of differential correlation until the findings have been validated in independent samples.

5. The future of Ryff's scales and the six-factor model

There is compelling empirical evidence of theoretical and/or measurement problems with Ryff's six-factor model of psychological well-being. Numerous bodies of survey data and alternative model specifications lead to the conclusion that four of the six RPWB factors are virtually indistinguishable. Standard methods of measuring RPWB are confounded by method effects that are at least as strong as several of the theoretical distinctions intended by the six-factor model. Moreover, available evidence of differential relationships between RPWB factors and their correlates is fully consistent with our findings about the factor structure of RPWB. These findings are bad news indeed for Ryff's six-factor model of psychological well-being.

Ryff and Singer's first objection to our paper is that we focus on "what PWB is *not*, rather than what it *is*" (p. 1103). We recognize that studies reporting bad news about widely implemented measures may be unwelcome, and they certainly can seem far less constructive than those which mostly affirm existing practice as correct. We, too, regard it as unfortunate that the evidence so plainly indicates the need to reconsider the measurement and meaning of psychological well-being. Again, we have no specific argument with Ryff's theory: There may well be a six-factor structure of psychological well-being, but the items Ryff has proposed to represent that structure fail to confirm it. The problem may lie in the instrumentation, in the theory, or in both.

As for practical recommendations for researchers working with the existing scales, we think that researchers should avoid RPWB indexes altogether and, instead, model the covariance structure of the items and their relationships with other variables. We believe

that Springer and Hauser (2006), along with the present text, provide ample evidence why that should be a preferred analytic strategy.

Where structural equation modeling is not an option, we would suggest either combining all of the items into a global well-being index or combining the four redundant subscales into one index and treating the other two dimensions separately. In future data collection, resources would be better allocated and statistical reliability improved by using fewer of the items pertaining to the highly redundant RPWB factors and adding more measures of autonomy and positive relations with others. Also, the proximity effects that we have identified could be attenuated in future studies by dispersing RPWB items among other items with similar agree-disagree response categories.

Researchers should be far more confident in their ability to reliably assess relationships between variables and global well-being than in its specific dimensions, especially with respect to the four subscales with the highest inter-factor correlations. Researchers should not take the fact that relationships between a variable and two dimensions of RPWB fall on different sides of the $p < 0.05$ line as being evidence of differential correlation, but should actually test whether correlations are significantly different from one another. Researchers looking to draw conclusions about relationships with specific subscales should regard replication in multiple samples as the evidentiary "gold standard," and thus they should seek to test hypotheses across multiple datasets whenever possible, e.g., as in Pudrovska et al. (2005). Researchers who work with large sample surveys might even consider disciplining analyses by conducting analyses first on a random half of the sample and, only after having obtained final results for that half, looking to see if the same pattern of results actually holds in the other half. Barring such evidence or analyses disciplined by statistical models like those in Fig. 1, researchers should anticipate that many differences in the pattern of results across subscales will later come out differently in independent samples.

These suggestions for research practice beg the main question, "How should psychological well-being be conceived and measured?" As we wrote in our original paper, Ryff's development of the six-factor model of psychological well-being is a valuable contribution to social psychological measurement. Moreover, the work has been highly influential: Two key papers—Ryff (1989) and Ryff and Keyes (1995)—have been cited in more than 500 published works. Thus, our analyses demonstrate a pressing need to rethink and recast current ideas about the structure of psychological well-being and/or about its measurement. This endeavor will require the integration of careful and critical theoretical, methodological, and empirical work.

The scientific endeavor is not well-served by the suggestion that closely interrogating the properties of measures and how they are affected by instrumentation practices—projects common to all scientific disciplines—is somehow spoilsport "methodological hammer[ing]" (p. 1116). The endeavor is even less well-served by the idea that the worthiness of responding to scientific challenges should depend less on their validity than on whether their authors meet arbitrary standards of adequate "substantive interest" (p. 1116). In this respect, Ryff and Singer's declaration that they will not engage in further scientific discourse about the measurement of psychological well-being may be the worst news yet for the six-factor model. We hope that other scientists interested in the conceptualization and measurement of psychological well-being will judge the merits of arguments rather than misjudge the motives of authors.

References

Cheng, S-T., Chan, A.C.M., 2005. Measuring psychological well-being in the Chinese. Personality and Individual Differences 38 (6), 1307–1316.

Clarke, P.J., Victor, M., Carol, D.R., Wheaton, B., 2001. Measuring psychological well-being in the Canadadian study of health and aging. International Psychogeriatrics 13 (1), 79–90.

DeVellis, R.F., 2003. Scale Development : Theory and Applications, second ed. Sage Publications, Thousand Oaks, CA.

Hauser, R.M., Goldberger, A.S., 1971. The treatment of unobservable variables in path analysis. In: Costner, H.L. (Ed.), Sociological Methodology. San Francisco, Jossey-Bass, pp. 81–117.

Hauser, R.M., Springer, K.W., Pudrovska, T., 2005. Temporal structures of psychological well-being: continuity or change. Presented at the 2005 Meetings of the Gerontological Society of America, Orlando, Florida.

Jöreskog, K.G., Dag Sörbom, 1996a. LISREL 8 User's Reference Guide, second ed. Scientific Software International, Chicago, IL.

Jöreskog, K.G., Dag Sörbom, 1996b. PRELIS 2 User's Reference Guide : A Program for Multivariate Data Screening and Data Summarization : A Preprocessor for LISREL, third ed. Scientific Software International, Chicago, IL.

Kafka, G.J., Kozma, A., 2002. The construct validity of ryff's scales of psychological well-being (SPWB) and their relationship to measures of subjective well-being. Social Indicators Research 57, 171–190.

Loehlin, J.C., 2004. Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis, 4th ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Pudrovska, T., Hauser, R.M., Springer, K.W., 2005. Does psychological well-being change with age? Presented at the 2005 Meetings of the Gerontological Society of America, Orlando, Florida.

Raftery, A.E., 1995. Bayesian model selection in social research. In: Marsden, P.V. (Ed.), Sociological Methodology. Blackwell, Basil, Cambridge, pp. 111–163.

Ryff, C.D., 1989. Happiness is everything, or is it? explorations on the meaning of psychological well-being. Journal of Personality and Social Psychology 57 (6), 1069–1081.

Ryff, C.D., Keyes, C.L., 1995. The structure of psychological well-being revisited. Journal of Personality and Social Psychology 69 (4), 719–727.

Ryff, C.R., Singer, B.H., 2006. Best news yet on the six-factor model of well-being. Social Science Research 35, 1102–1118.

Springer, K.W., Hauser, R.M., 2006. An assessment of the construct validity of Ryff's scales of psychological well-being: method, mode, and measurement effects. Social Science Research 35 (4), 1079–1101.

Van Dierendonck, D., 2004. The construct validity of Ryff's scales of psychological well-being and its extension with spiritual well-being. Personality and Individual Differences 36 (3), 629–643.

Weakliem, D., 1999. A Critique of the Bayesian information criterion in model selection. Sociological Methods and Research 27 (3), 359–397.