The (In)validity of Self-Construal Scales Revisited

TIMOTHY R. LEVINE
MARY JIANG BRESNAHAN
HEE SUN PARK
Michigan State University
MARIA KNIGHT LAPINSKI
Western Michigan University
TAI SIK LEE
DONG WOOK LEE
Hanyang University

This paper offers a critical rejoinder in the discussion concerning the viability of self-construal scales. It is our contention that the existing data reported in our previous article (Levine et al., 2003) and elsewhere are more than sufficient to justify the conclusion that self-report scales purporting to measure interdependent and independent self-construals as 2 orthogonal constructs lack validity. The arguments to the contrary offered by Gudykunst and Lee (2003) and Kim and Raja (2003) are disputed. Additional data (N = 1,013) show that neither age, occupation, sample size, standardization, nor a 2nd-order structure can account for the problems we documented previously. Although we see potential utility in the self-construal construct, we believe that the 3 primary scales fail to meet reasonable and accepted social scientific standards. Difficult conceptual problems will need to be solved prior to the development of new and improved measures.

Measurement validity cannot be achieved by assertion. It must be empirically demonstrated. In the absence of evidence consistent with validity, there is no presumption of validity. If and when data accumulate that are consistent with a scale’s validity, our confidence in a scale gradually increases. However, when data are inconsistent with what is expected given extant conceptualization or theory, our confidence in the measure ought to decline accordingly. If the data appear to be mixed, a high degree of confidence is not warranted and validity is open to debate.

Timothy R. Levine (Ph.D., Michigan State University, 1992) and Mary Jiang Bresnahan (Ph.D., University of Michigan, 1985) are professors in the Department of Communication at Michigan State University, where Hee Sun Park (M.A., University of Hawaii, 1998) is an assistant professor. Maria Knight Lapinski (Ph.D., Michigan State University, 2000) is an assistant professor in the Department of Communication at Western Michigan University. Tai Sik Lee is a professor and Dong Wook Lee is a doctoral student in the Department of Civil and Environmental Engineering at Hanyang University in Ansan, South Korea.
The demonstration of validity requires a variety of forms of evidence. First, for a set of items measuring a single construct, a minimum requirement is that a stable and unidimensional structure be shown to exist. That is, items that are summed or averaged as a measure of a single construct must act as if they are measuring a single construct. If more than one thing is measured (i.e., the scale is multidimensional), then the measure is confounded and the meaning of summed or averaged scores is ambiguous. In the case of interdependent and independent self-construal scales, the items, at minimum, must be shown to measure two and only two constructs, with all interdependence items measuring one construct and all independence items measuring a second construct orthogonal to the first. Further, because these scales are specifically intended for cross-cultural use, their structure must also be reasonably invariant across the range of applications for which the scales are used. Only after dimensionality has been adequately documented can reliability and the scale’s association with other constructs of interest be meaningfully interpreted. Then, additional evidence for validity may include placement of the scale within a nomothetic network (i.e., correlating the scale with other constructs), establishing convergent validity (i.e., showing that alternative measures of the same construct function similarly), and establishing predictive utility.

We contend that self-construal scales lack validity for a variety of reasons. First, our meta-analysis showed massive heterogeneity of effects, thus suggesting powerful artifacts. Oyserman, Coon, and Kemmelmeier (2002) have independently published findings that are strikingly consistent with our own. Second, priming experiments showed that self-construal scales are insensitive to priming whereas previous research consistently demonstrates that alternative measures of self-construals are affected by priming. This constitutes a failure of convergent validity because alternative measures of the same construct should behave similarly. Finally and most critically, the interdependent and independent self-construal scales that we examined were neither stable, unidimensional, nor invariant. The failures were neither subtle nor ambiguous. Additionally, these findings are not unique to our study. Separately, each of these findings is sufficient to raise serious concerns. Together, it is not hyperbole to say that our findings provide convincing, and we believe indisputable, evidence of validity problems.

COMMENTS ON GUDYKUNST AND LEE (2003)

Gudykunst and Lee (2003) provide professional, thoughtful, and constructive comments on self-construals, and, on many issues, we are in agreement. We disagree, however, on the core issue of the viability of the
two-dimensional self-construal scales. Gudykunst and Lee argued that the results of our meta-analysis and priming studies are compatible with the validity of self-construals. Although they grant that the findings of our measurement analyses might pose problems for self-construal scales, they speculate that the scales might be second-order valid. Finally, Gudykunst and Lee assert that 50 studies have produced theoretically consistent results, and this would be impossible unless the scales were valid. We provide logical or data-based refutations of each of these arguments below.

The Meta-Analysis

Gudykunst and Lee (2003) attribute the apparent failures of the self-construal scales in the meta-analysis to the use of student and Japanese samples. As they state:

> If the self construal scores do not fit the “expected” patterns, it does not necessarily mean that the scales are invalid. It probably means that the samples are not representative of cultural individualism-collectivism. (p. 256)

On one level, this argument is perfectly reasonable, and we do not disagree that evidence showing age effects on interdependent self-construals exists. For example, Rao, Singhal, Ren, and Zhang (2001) recently reported just such an effect (although the effect size was a very small $r = .12$). On the other hand, an obvious problem with this argument as a post hoc explanation of our findings, is that it is circular and lacks falsifiably. It dismisses all conflicting findings by definition. That is, if the scores come out right, the sample is representative. Proof that the sample is representative is that the scores came out right.

We nevertheless believe this issue is more constructively treated as an empirical question. Therefore, we collected additional data ($N = 1,013$). The participants included 295 undergraduate students at Michigan State University (62.9% female, ranging in age from 18 to 44, $M = 20.7$, $SD = 2.2$), 311 Korean undergraduate students (75% male, ranging in age from 18 to 28, $M = 21.8$, $SD = 2.3$), 48 Korean graduate students (all female, ranging in age from 23 to 34, $M = 26.0$, $SD = 2.9$), and 354 Korean workers (67.2% male, ranging in age from 20 to 49, $M = 31.9$, $SD = 5.5$). The nonstudent Korean participants worked in the construction industry (52%), education-related jobs (mostly kindergartens, 16%), government agencies (12%), telecommunications (7%), and a variety other occupations (information technology, automobile manufacturing, landscaping, travel, and finance). All participants completed the self-construal items (with 7-point response formats) provided by Gudykunst and Lee (2003) as well as demographic questions.
Consistent with self-construal predictions and previous findings, Americans scored significantly higher on independent self-construals ($M = 5.97$) than did Koreans ($M = 5.19$), $F(1, 1011) = 424.02, p < .001, \eta^2 = .30, r = .55$. Wrong direction findings, however, were observed for interdependent self-construals. Americans scored significantly higher ($M = 5.24$) than Koreans ($M = 5.03$), $F(1, 1011) = 27.90, p < .001, \eta^2 = .03, r = .16$. Further, Koreans scored significantly higher on independent than interdependent self-construal, $t(717) = -6.01, p < .001$.

Similar to Rao et al. (2001), there was a small positive correlation between scores on interdependent self-construals and age within the Korean samples, $r(695) = .178, p < .001$. No such correlation was observed for Americans, $r(293) = -.034, p = ns$. Although Korean workers ($M = 5.19$) were more interdependent than Korean graduate students ($M = 5.08$), who were more interdependent than Korean undergraduates ($M = 4.95$), the mean for the Korean workers was still below the mean for American undergraduates ($M = 5.24$). These data suggest that neither age nor occupation explain the problems observed in our meta-analysis.

The logic of Gudykunst and Lee’s (2003) argument holds that because the Koreans in our latest sample scored more highly on independence than interdependence and scored lower on interdependent self-construal than American students, they, as a sample, must not reflect cultural-level collectivism. Circular reasoning aside, is it not plausible that the Koreans in our sample do indeed embrace Korean culture and that either the scale is not measuring what it is intended to measure or that the theory behind it not consistent with the data?

**Priming**

Gudykunst and Lee (2003) acknowledge that a substantial and consistent literature exists showing that priming influences the self-construals that individuals “activate.” They assert, however, that scores on self-construal scales measure the self-construals that are formed as individuals are socialized, not situationally activated self-construals. They therefore conclude that scores should be “relatively stable,” that “priming only influences the self construals individuals activate, not their scores on the scales,” (p. 266) and that the priming findings are therefore perfectly consistent with the validity of the scales.

This argument requires a tenuous logical leap. It is difficult to imagine how self-report scales could possibly provide a valid measure of stable self-concepts and simultaneously remain totally impervious to the “activated” self. Is it not more plausible to predict that the activated self should be reflected in scale responses and the fact that they are not reflects a problem?

The second problem with Gudykunst and Lee’s (2003) argument is that other measures of self-construal (e.g., the 20-questions test) are situation
(Kanagawa, Cross, & Markus, 2001) and priming (e.g., Gardner, Gabriel, & Lee, 1999) sensitive. This apparent lack of convergent validity would seem to suggest that either one or the other measure is problematic or that they tap different constructs. Further elaboration is provided in the response to Kim and Raja (2003) below.

Measurement Studies

Gudykunst and Lee (2003) attempt to reconcile our findings of lack of fit and of more than two factors with the possibility of two second-order factors. As they note, multiple “factors could easily form two factors in a second-order factor analysis” (p. 261). In theory, Gudykunst and Lee are absolutely correct. Our findings of more than two factors would not be inconsistent with the validity of the self-construal scales if two higher-order factors (one for independence and the other reflecting interdependence) could be demonstrated.

Given our experience with both second-order analyses (e.g., Levine, 1998; Levine & McCroskey, 1990) and with the self-construals scales in question (Levine et al., 2003), however, we think the probability of this possibility is vanishingly small. The interitem correlations are neither large enough, nor patterned in such a way as to suggest a second-order structure. The evidence presented in our original measurement studies is also inconsistent with a second-order model. If a second-order structure was functioning, we would expect the lack of fit to be more evident in the internal consistency matrices than in the parallelism tests. The data we presented in the measurement studies uniformly and reliably showed greater problems with parallelism than internal consistency. This, too, is inconsistent with the existence of second-order factors.

Nevertheless, we again treated Gudykunst and Lee’s (2003) speculation as an empirical question. We first tested the original two-factor model with our new data. Gudykunst and Lee imply that one reason for the lack of fit in our previous analysis was our decision to conduct within-nation analyses (i.e., that their scale should be tested with pan-cultural analyses). A pan-cultural (N = 1,013) confirmatory factor analysis (CFA) using PACKAGE computer software and specifying the a priori two-factor solution with the new data showed a profound lack of fit. Statistically significant deviations (at \( p < .05 \)) were observed in 43% of the correlations in the internal consistency test for independent self-construal (with 27% significant at \( p < .01 \)), 49% of the correlations in the internal consistency test for interdependent self-construal (with 36% significant at \( p < .01 \)), and in 55% (41% at \( p < .01 \)) of the correlations in the test of parallelism. Following Gudykunst and Lee, we next standardized scores within individuals, then within culture, and retested the model. Again the two-factor model failed. A lack of fit was observed within both the Korean and U.S. samples.
These data show that the lack of fit reported in our previous paper replicates with the Gudykunst and Lee items, a larger sample, with and without standardization, and with both pan-cultural and within-country analyses.

We investigated the possibility of a second-order factor structure by trying to isolate a stable first-order structure that could be subsequently tested for second-order bidimensionality. An iterative series of exploratory pan-cultural maximum likelihood and principal axis factor analyses with oblique rotations suggested the possibility of four factors (independent self-construal, self-reliance, group cohesion, and a consulting communication style) after many weak items were deleted. We could not, however, find a first-order solution that (a) held in both the Korean and U.S. samples, (b) provided acceptable fit with CFA, or (c) retained a sufficient number of items and factors to allow second-order factor analysis. We therefore conclude that a second-order structure cannot explain the lack of fit for the a priori two-factor solution, and that the problems with the scale cannot be easily solved by deleting items.

A final comment we have regarding Gudykunst and Lee’s (2003) response to our measurement studies is in regard to note 9, where several plausible reconceptualizations of self-construals are listed. The very existence of multiple, readily available, viable theoretical alternatives suggests that self-construals are on shaky theoretical footing. If sound theory existed, surely scholars would know which, if any, of these alternatives should guide the measurement. A thoughtful read of Gudykunst and Lee (2003) as well as Kim and Raja (2003) suggests that “a clear theory” has not yet been developed. Thus, the measurement problems are not surprising.

The 50 Previous Studies and Construct Validity

Perhaps Gudykunst and Lee’s (2003) most persuasive argument is that:

Evidence for the construct validity of the self construal scales . . . indicates that there are theoretically consistent findings across approximately 50 studies using the three scales. This would not be possible if there were major problems with the two dimensional model of self construals or the scales used to measure them. (p. 253)

What Gudykunst and Lee (2003) refer to as “construct validity” is actually only one approach to construct validation—one that we will call the “nomological network approach” developed by Cronbach and Meehl (1955). The nomological network approach, although probably the most widely known, is not the only, or even the strongest, approach to construct validity. Thus, when Gudykunst and Lee assert that self-construal scales exhibit construct validity, they are only addressing one aspect of construct validity. Nevertheless, this is a line of argument we anticipated.
First, Gudykunst and Lee’s (2003) argument exploits what Abelson (1995) calls the “hocus focus trick.” Hocus focus involves pointing out supportive evidence while ignoring nonsupportive evidence. It is true that there are studies that appear highly and consistently supportive (e.g., Ellis & Wittenbaum, 2000; Oetzel et al., 2001). There are also some apparent disconfirmations (e.g., Park, 2001). However, much of the published work on self-construals is very much a mixed bag. Many studies test multiple hypotheses where some hypotheses are supported while others aren’t (e.g., Kim, Aune, Hunter, Kim, & Kim, 2001; Kim et al., 1996; Oetzel, 1998). Also, many effects that are statistically significant reflect very small effect sizes. So, the data aren’t nearly as neat or as consistent as Gudykunst and Lee assert.

Second, as Meehl (1978, 1986) points out, the interpretation of significance-test counts is ambiguous. For example, studies reporting significant results are more likely to be published than those with nonsignificant results and studies with supportive results are more likely to survive the review process than those that contradict prevailing wisdom. These processes (and others) can result in an inflation of apparent support (see Meehl, 1978, 1986).

Third, an argument made in our initial article is that some apparent evidence for construct validity might be an artifact of the multidimensional nature of self-construal items. Gudykunst and Lee (2003) disagree, arguing that there is “no evidence that the correlations are due to the self construal scales containing items related to the concepts with which they are correlated” (p. 265). We respectfully disagree. For example, several items in Gudykunst and Lee’s appendix explicitly reference consulting with others. Is it not possible that items asking people about consulting with others measure a communication style, rather than, or in addition to, a self-concept?

Much more obvious examples, however, can be seen in Professor Kim’s research. For example, Gudykunst and Lee (2003) list Kim et al. (2001) and Kim et al. (1996) as providing evidence for construct validity. Kim et al. (2001) report that independent self-construal is negatively associated with communication apprehension (CA), while interdependent self-construal is uncorrelated with CA. Yet, the Kim and Leung independent self-construal scale contains items like “I voice my opinions” and “speaking up . . . is not a problem for me.” These read more like communication apprehension items than self-concept items. As another example, Kim et al. (1996) found that scores on interdependent self-construal were positively associated with the conversational constraint “concern for others’ feelings.” Kim and Leung’s scale contains items like “I conceal my negative emotions so I won’t cause unhappiness” and “I watch my words so I won’t offend anyone.” On face, these items sound more like concern for others’ feelings than self-concept. Given obvious examples like these,
we are puzzled that Gudykunst and Lee can claim that this is not a problem. It is indeed plausible that much of the apparent evidence for construct validity is artifactual.

In summary, Gudykunst and Lee (2003) argue that:

The evidence consistently supports the construct validity of the three self construal scales. This would not be possible if there are “severe” or “fatal” (or even major) flaws with the conceptualization of self construals or the scales designed to measure them. (p. 264)

Our point is that the published evidence is not consistent, and, even if it was, it would still be possible for major flaws to exist. Dimensionality is a prerequisite for correlations with other variables, not vice versa. Gudykunst and Lee’s (2003) argument rests on backward logic.

This said, we are not so cynical as to think or argue that there is nothing there. Self-construal scales do indeed seem to be measuring something (and we suspect, many things). Instead, we simply note that the results of the 50 studies would still be possible if our thesis is true, and that the fact that the scales correlate significantly with outcome measures does not provide proof of validity convincing enough to override other evidence to the contrary.

Conclusion

Gudykunst and Lee (2003) offer counter arguments for most of our data and arguments. Taken separately, most of their arguments sound reasonable. If (a) the samples in the meta-analysis were not representative, (b) if sampling could reasonably explain the heterogeneity of findings in the meta-analysis, (c) if activated self-construals should not affect scores on the scales, (d) if there is really a second-order structure to the scales, (e) if all 50 studies really found consistent support, and (f) if there is not a substantial body of unpublished failures, then a very strong and legitimate case can be made for the validity of the scales. We, however, believe that a much more parsimonious and plausible account of the data is simply that validity problems exist (although we admit to some sympathy with Gudykunst and Lee’s construct validity argument).

COMMENTS ON KIM AND RAJA

Professor Kim has been an innovative and prolific self-construal scholar, and her work correlating self-construals scores with a wide variety of communication scales has been highly influential. She has also formally tested self-construal mediation models with path analysis (e.g., Kim et
al., 1996; Kim et al., 2000; Kim et al., 2001) and is one of the few self-construal researchers to report CFAs in some of her work (e.g., Kim et al., 2001; Kim, Shin, & Cai, 1998; Tasaki, Kim, & Miller, 1999).

Kim and Raja (2003) would have the reader believe that we are unscientific, illogical, and possibly unethical. They charge that “due to ethnic stereotyping and other erroneous assumptions, as well as faulty logic and unsound methodology, Levine et al.’s (2003) conclusions on crucial aspects of validity are untenable” (Kim & Raja, pp. 276–277). We, of course, disagree with this characterization and think that both our arguments and data have merit.

We strongly suspect that readers familiar with measurement validity will not see our criteria as especially unreasonable or as overly stringent. For example, most would agree that a scale that is supposed to tap two factors should reflect two factors. Second, we believe that those readers who have read Professor Kim’s work will find that we have taken care to cite her work accurately and that it is not us who are setting up and knocking down straw persons. Third, we believe that part of being committed to social science is the presentation of, and reliance on, objective data. Finally, we would maintain that an ad hominem attack does not pass as evidence in scientific debate and that a sound argument by analogy requires that the analogy is relevant and appropriately applied to the claim.

In our view, Kim and Raja (2003) err in the very first sentence of their abstract where they imply that we proclaimed the death of self-construal research. To the contrary, we see much potential in the construct. Instead, it is our much more modest claim that three commonly used self-construal scales, including one constructed and used by Professor Kim, are not valid.

Consider also this example of Kim and Raja’s (2003) logic. On page 276 they state that “Carefully developed concepts are, in turn, a prerequisite for meaningful discussions of measurement validity.” On page 286, however, they acknowledge that “we are in no position to give a clear picture of the structure and conceptualizations of the independent and interdependent selves.” If both of these statements are true, then it follows that Professor Kim is in “no position” to construct, much less defend, her own measure and that the position we have taken is correct by default, that is, the scale must be questionable if its author does not know how the construct should be conceptualized or how the scale should factor.

Our Meta-Analysis and Cultural Lumping

Kim and Raja (2003) accuse us of “crude ethnic stereotyping” that is “unscientific.” Specifically, they seem to take particular offense at our use of the terms “Western” and “Asian.” It was not our intent to offend, and we meant only to employ terminology common in the literature (e.g., Kim et al., 1998; Kim et al., 2001; Markus & Kitayama, 1991), including
Professor Kim’s recent book (Kim, 2002). We purposely did not use the terminology adopted by Gudykunst and Lee (2003; i.e., cultural level individualism and collectivism) because our reading of literature suggests that the problems we observed with self-construals are equally evident in individualism and collectivism (e.g., see Fiske, 2002; Oyserman et al., 2002).

More importantly, however, Kim and Raja (2003) seem to miss our point. We are not arguing that people from different countries have systematically different self-construals. We oppose that view. Kim and Raja criticize us for a position that is not our own.

Our Hypotheses 1–4 were derived from the literature, including Professor Kim’s own work. Markus and Kitayama (1991) originally invoked the term “self-construals” to explain Western, non-Western differences (their words, not ours). Gudykunst and his associates (Gudykunst et al., 1996; Gudykunst & Lee, 2003) and Kim (Kim, 1996; Kim et al., 2000; Kim et al., 2001; Kim et al., 2002), claim that self-construal is a mediator of cultural effects. Asserting that self-construal is a mediator posits, by definition, that culture is the proximate cause of self-construals and therefore specifies cross-cultural differences. In fact, one real irony of Professor Kim’s (Kim & Raja, 2003) criticism of our hypotheses is that she has repeatedly hypothesized, tested, and claimed support for the very relationship that she criticizes us for testing (e.g., see Kim, 1996; Kim et al., 2000; Kim et al., 2001; Kim et al., 2002).

Both Gudykunst and Lee (2003) as well as Kim and Raja (2003), stress that self-construals are trait-like variables that are intended to explain within-culture variance. Kim and Raja further imply that it is improper to expect systematic ethnic and national differences. If this is true, then does not the link between culture and self-construals become lost?

Priming

Kim and Raja (2003) are critical of our interpretation of our priming studies. First, they provide extensive critique of one of our three rival hypotheses. We think that the hypothesis in question follows from the self-construal literature (e.g., Gaertner, Sedikides, Vevea, & Iuzzini, 2002; Haberstroh, Oyserman, Schwarz, Kuhnen, & Ji, 2001; Kanagawa et al., 2001; Markus & Kitayama, 1991; Singelis, 1994). Kim and Raja disagree. We see this as a legitimate disagreement over the reading of the literature and not as “a logical fallacy.”

Kim and Raja (2003) argued that “the original purpose of the self-construal scale was to tap the stable and trait-like aspects of self-representations” (p. 281). To the extent that this is true and to the extent that this is consistent with the conceptual definition of self-construals, we would agree completely that the scores should be impervious to priming. We, however, doubt that all self-construal researchers would agree that self-construals are trait-like.9
Finally, Kim and Raja (2003) speculate that the 20-questions test measures the dynamic aspect of self-construals whereas the scales under examination here measure the trait-like aspects of self-construals. Therefore, the fact that they fail to converge is not grounds for criticizing the scales. Kim and Raja make a legitimate point. We do not disagree, and we see this as very much consistent with our data and conclusions. We would note, however, that it follows then that the 20-questions method as compared to the self-construal scales must therefore measure conceptually and empirically distinct constructs. Calling these different constructs by the same name invites theoretical confusion. But, given Kim and Raja’s admission that they “are in no position to give a clear picture of the structure and conceptualizations of the independent and interdependent selves” (p. 286), we question how they would know what to hypothesize or how they can be in a position to assert that self-construals are trait-like.10

Factor-Analyses

Besides confusing exploratory factor analysis with principle components analysis, Kim and Raja (2003) criticize our CFAs because we specified two factors. They argue that it is unreasonable to expect the self-construal scales under consideration to reflect two factors. Our response is as follows. Professor Kim has constructed her own measures of self-construals. She designed them specifically to yield two scores, one for independent self-construal and one for interdependent self-construal. In fact, she has recently and explicitly stated in print that she “conceptualized Markus and Kitayama’s (1991) self-construals of independence and interdependence as two orthogonal dimensions of the self-concept” (Kim et al., 2001, p. 395). In every published study using a self-construal scale (that we have been able to locate), she has scored the scales as exactly two factors, never more, never less. Further, in every published study where Professor Kim has reported a CFA of her scales, she has specified exactly two factors. She has also stated that in her own research, “confirmatory factor analysis was used to validate the factor structure of the independent and interdependent scales” (Kim, Shin, & Cai, 1998, p. 56). In her recent book, Professor Kim (2002) endorsed the two-factor model and suggested that existing evidence supports the two-factor model. It seems completely reasonable for us to conclude upon reading Professor Kim’s own published research that (a) her scale was indeed intended to tap two factors and that (b) she thinks that CFA is an acceptable method to test validity.

We believe that it is neither unscientific, nor illogical, nor setting up a straw person to seek to replicate another scholar’s findings. If Professor Kim honestly believes that it is premature to specify the number of fac-
tors for her scale, then on what basis did she publish a number of studies predicated on the existence of exactly two factors? Moreover, if Professor Kim believes that she is justified in scoring her scale as two factors, then how can she possibly justify her criticism of us for testing a two-factor model unless she rejects the desirability of independent replication by other scholars? If this is the case, then it is not us who are unscientific.

It is noteworthy that Professor Kim has reported using CFA on her self-construal scales on at least three occasions (Kim et al., 1998; Kim et al., 2001; Tasaki et al., 1999). In two cases, these data were presented in such a way as to imply that the two-factor model was confirmed (Kim et al., 1998; Tasaki et al., 1999). In the other case (Kim et al., 2001, pp 395–396), Professor Kim reported results that constitute an unequivocally unacceptable fit.11 In that article, she nevertheless treated her scales as two factors, tested a causal model using those two scores, and drew substantive conclusions about a causal model, including nations as culture and the two dimensions of self-construals as partial mediators of nation-level culture. Professor Kim therefore either knows, or should know, that her scales have severe validity problems.

Summary

Unlike Gudykunst and Lee (2003), Kim and Raja (2003) do not defend the two-factor model of self-construals, or dispute our findings. Instead, they chose to attack our questions, arguing forcefully that we were wrong to test our hypotheses or even to pose them. Kim and Raja reject the prediction of national or ethnic differences in self-construals as a validity criteria (even though they have not infrequently made such predictions themselves). When self-construals fail to show the anticipated differences, Kim and Raja not only conclude that it was wrong to expect differences in the first place, but further argue that the lack of differences actually highlights the importance of the construct. They reject convergent validity arguments involving alternative measures of self-construals. When the two measures function differently, they again interpret the findings as consistent with the construct. They even reject the specification of an a priori factor structure (despite the fact that Professor Kim has published a considerable body of work specifying the same two-factor a priori structure they criticize us for). This, we believe, renders Kim and Raja’s position unfalsifiable.

CONCLUSIONS

In social scientific debate, the critic is at a natural advantage. It is a much easier intellectual task to criticize existing scholarship than it is to
create new knowledge. Critics also have an inherent logical advantage given the relative power of falsification over verification arguments. Finally, critics choose what to critique, thereby having the luxury of selecting positions that are relatively easy to argue. With these advantages, we believe, comes obligation. Critics such as ourselves should be held to a higher logical and evidentiary standard.

We took the critic’s obligation seriously. A wide array of original data consistent with our arguments was presented in our original article. Extra effort was taken to replicate our findings multiple times. We also documented published findings by other scholars that are consistent with our own results. We did not pursue publication of these findings until we were very confident in our conclusions.

Gudykunst and Lee (2003) provided rival explanations for many of our original findings. We have presented new data in this rejoinder testing Gudykunst and Lee’s sampling and second-order arguments. The data were not consistent with these rival accounts. Instead, the data continue to provide a coherent picture that is inconsistent with the viability of self-construal scales.

What’s Next

The one (and maybe the only) thing that all three sets of authors in this colloquy agree upon fully is that a clear and sound conceptual foundation is an absolutely necessary prerequisite for valid measurement development. Content in both the Gudykunst and Lee (2003) as well as the Kim and Raja (2003) replies, lead us to conclude that the conceptualization of self-construals, at least among communication scholars, is nowhere near as clear as some of the articles in the literature imply. Therefore, if this debate is to have a constructive outcome, some difficult conceptual challenges will need to be met.

We listed a number of conceptual challenges in our initial article, and we continue to believe that these issues need to be resolved before new measurement development takes place. One issue we raised was that of autonomy versus stability. Both Gudykunst and Lee (2003) as well as Kim and Raja (2003) take the view that self-construals should be considered trait-like. Whereas we too suspect that self-concept flexibility is different from self-in-relation-to-others, we hope that communication research will not be exclusively restricted to either trait-like models of the self or to Likert-scale methods. For example, the approach of Kanagawa et al. (2001) is an alternative worthy of consideration.

Gudykunst and Lee (2003) raise an important concern about orthogonality. If scholars rethink the number of dimensions of self-construals, they may also want to rethink how those dimensions are interrelated. Constructs can be distinct without being orthogonal. We fail to see what
is theoretically gained by the specification of orthogonality and perhaps removing the orthogonality requirement would be fruitful.

We will also raise a new conceptual issue. Gaertner et al. (2002) as well as Sedikides, Gaertner, and Toguchi (in press), argue for the primacy of the individual self. In their view, the independent self may be culturally invariant whereas a collectivist self is more likely shaped by norms and culture. Alternatively, Andersen and Chen (2002) argue that the relational selves may be universal and that “regardless of culture or gender, the self and personality should be in part defined by a person’s idiosyncratic set of relational selves” (p. 637). These ideas had occurred to us and they might be worthy of further consideration. Given these alternative perspectives, coupled with Kim and Raja’s (2003) new assertion that we should not even hypothesize ethnic or national differences, a major challenge for intercultural communication scholars is to keep culture from getting completely lost when conceptualizing self-construals.

In any case, once a satisfactory conceptual explication of self-construals has been achieved, those attempting to construct new or revised scales will face a number of additional challenges. In addition to the validation issues facing all measures, self-construal scales will need to deal with issues such as the reference-group effect (Heine, Lehman, Peng, & Greenholtz, 2002) and response alternative effects (Ji, Schwarz & Nisbett, 2000). Further, if communication scholars are committed to the idea of individual-level culture, multilevel modeling may be desirable (e.g., see VanDeVijver & Poortinga, 2002). In short, we think that progress in this area will require advances in both conceptual and methodological sophistication.

Final Comments

We believe that the data is unequivocal and that current self-construal scales lack validity. That said, we will reiterate that we are not arguing for a moratorium on self-construal research, and we hope this debate has a constructive outcome. For self-construal research to progress, however, a clearer conceptualization and better measurement options are needed. To the extent that self-construals are valuable constructs, they deserve adequate measurement. Abelson’s (1995) eighth law states that criticism is the mother of methodology. We sincerely hope this is true for cross-cultural communication research.

NOTES

1. This is a crucial point. Invalid scales can be reliable, and reliability, by itself, is not sufficient evidence for validity. In fact, as noted in our initial article, it is mathematically
possible for certain measurement confounds to inflate reliability. A scale must be unidimensional (at some level) before it is possible to meaningfully interpret sums, averages, reliabilities, or associations with other variables. If we are right and neither the independent nor the interdependent self-construal scale is unidimensional, then other evidence that presupposes unidimensionality, such as reliability or construct validity (that relies on correlations with other measures), is immaterial, as are the results of our meta-analysis and priming studies. That is, the findings of the measurement studies alone are sufficient to justify our conclusion.

2. We see the issue of the heterogeneity of effects in the meta-analysis as the most important finding from our meta-analysis and as constituting strong evidence consistent with our argument. We showed that different studies testing for the same differences produced findings that were off by as much as 10 standard errors. Yet, neither Gudykunst and Lee (2003) nor Kim and Raja (2003) comment on this part of our results. Consequently, they fail to explain how a valid measure could possibly produce such highly variable results. Both Gudykunst’s student sample argument and Kim and Raja’s false dichotomy argument can be argued to predict the opposite of our finding—homogeneity of effects. Thus, neither response is consistent with the most important feature of our meta-analytic data. A data-based explanation of the heterogeneity consistent with our own view is presented in our next footnote.

3. Similar to our own meta-analysis, Oyserman et al. (2002) report massive heterogeneity in the individualism-collectivism literature. Although they were not able to explain all the heterogeneity of effects, they did find that item content was a substantial moderator. This means that individual items functioned differently in their meta-analysis. This is conceptually and empirically equivalent to a significant lack of parallelism in confirmatory factor analysis (CFA) and constitutes evidence of multidimensionality and a lack of a two-dimensional first- or second-order structure. Therefore, Oyserman et al.’s finding that item content is a strong moderator effectively shows how the problems with scale dimensionality reported in our measurement studies can produce the heterogeneous pattern of results we observed in our meta-analysis, as well as why the heterogeneity in the meta-analysis is an especially problematic finding for those who would defend self-construal scales.

4. The observed fit of the data to the a priori measurement model was, in every case, extremely poor. Nevertheless, Gudykunst and Lee (2003), citing debatable rules of thumb about sample size in factor analysis, argue that the lack of fit might be attributable to our sample sizes (N = 121, 223, 230, 323, 214, 206, 126, 204, 148, 141, and 150) and that larger samples might produce data more consistent with the validity of the scales. This argument is not statistically sound and reflects a misunderstanding of CFA. Sample sizes cannot provide a reasonable explanation for the lack of fit in our studies. First, the fit was so far off that we can rule out sampling error with a high degree of confidence (i.e., p < .0001). In fact, one of the most attractive advantages of CFA over typical EFA is that significance tests can, and are, used to rule out sampling error. Second, in confirmatory statistics (including our CFAs, but unlike traditional null hypothesis testing), the greater statistical power that comes with larger samples makes fit more difficult, not easier, to achieve. That is, with larger, more stable samples, ever smaller deviations from the model become statistically significant, making fit less probable. That numerous and highly significant deviations were observed even in our smallest sample provides testimony to how truly poor the observed fit was.

5. Since the writing of our original paper, we have learned of four additional papers testing the fit of the two-dimensional model. Kim et al. (2001) used Hunter’s CFA to test the internal consistency of an orthogonal two-factor self-construal scale. Even using very liberal fit criteria, numerous fit problems were observed in the independence scale. Rao, Singal, Ren, and Zhang’s (2001) data suggest nontrivial problems with the Singelis scale in China. Only 11 of 24 items were retained across their total sample. More problematic still, the scale factored differently for men and women, for people of different age groups, and for urban and rural groups, suggesting unacceptable instability. Similarly, Hsu (2002) used the Singelis scale in Taiwan and retained only 14 of 24 items. By contrast, however, Kwan, Bond, and
Singelis (1997) report that the Singelis scale provided an acceptable fit and was reasonably invariant in samples from the U.S. and Hong Kong. Thus, three out of the four additional studies we located suggest substantial problems similar to those we identified.

6. Professor Kim’s (Kim & Raja, 2003) claim that she “carefully removed those communication items from the previous self-construal scales (e.g., Singelis scale) in our research” (p. 285) is difficult to reconcile with the presence of the “voice my opinion” and “speaking up” items in her scale.

7. Kim et al.’s (1996) “concern for others feeling conversational constraint” items included “In this situation, I feel it is very important to avoid hurting other’s feelings” and “In this situation, being considerate towards other’s feelings is a major concern for me” (p. 40).

8. From our perspective, various claims made by Kim and Raja (2003) seem to exemplify what might be characterized as “meta-straw-persons.” That is, they make erroneous claims about the erroneousness of our claims. Put differently, their charge that we are making straw persons is, itself, a straw person.

9. The idea is that the interdependent self is flexible was made explicit in Markus and Kitayama (1991). As they noted, “an interdependent self . . . changes structure with the nature of the particular social context. Within each particular social situation, the self can be differentially instantiated” (p. 227). Similarly, Singelis (1994) defined interdependent self-construal “as a flexible, variable self.” (p. 581). Given that Singelis’s scale was historically first and that he endorses the idea that the interdependent self-construal is flexible, Kim and Raja’s (2003) claim that “the original purpose of the self-construal scale was to tap the stable and trait-like aspects of self-representations” (p. 281) would seem to be objectively false.

10. A Kim and Raja (2003) criticism of our priming studies is that our use of Midwestern U.S. participants renders our three experiments “unsound methodologically.” Previous priming studies conducted in the U.S., but using different measures, have found that priming affects self-construals. A critical part of our argument involved comparing our experiments to others, where the only substantive difference was how self-construals were measured. Given this, our selection of participants was justified.

11. Kim et al. (2001) reported that fully 30% of the deviations in an internal consistency test were significant at $p < .01$ (results are not reported at $p < .05$ and tests of parallelism were not reported). Additionally, they commented that “it is necessary to be cautious about the internal consistency of the independent self-construal scale” (p. 396).

REFERENCES


