Adjusting Bilingual Ratings by Retest Reliability Improves Estimation of Translation Quality

Dustin Wood

University of Alabama

Lin Qiu* and Jiahui Lu

Nanyang Technological University

Han Lin

Institute of High Performance Computing, Agency for Science, Technology and Research

William Tov

Singapore Management University

Citation:

Wood, D., Qiu, L., Lu, J., Lin, H., & Tov, W. (2018). Adjusting bilingual ratings by retest reliability improves estimation of translation quality. *Journal of Cross-Cultural Psychology, 49*, 1325-1339. doi: 10.1177/0022022118757914

This paper is not the copy of record and may not exactly replicate the authoritative document. The authoritative document is published in Journal of Cross-Cultural Psychology (https://doi.org/10.1177/0022022118789773), a Sage Publication.

Corresponding author*:

Lin Qiu, Division of Psychology, School of Social Sciences, Nanyang Technological University, 14 Nanyang drive, Singapore, 637332

Phone: +65 6513-2250

Fax: +65 6795-5797

Email: linqiu@ntu.edu.sg

ABSTRACT

The quality of cross-language scale translations is often explored by having bilingual participants complete the scale in both languages and then correlating their scores. However, low cross-language correlations can be observed due to score unreliability rather than due to poor scale translation. McCrae, Yik, Trapnell, Bond, & Paulhus (1998) suggested that a better indicator of translation quality can be formed by dividing the raw cross-language correlation by the same-language retest correlations over a similar measurement interval. Here, we illustrate how this method can be extended to evaluate the translation quality of individual items. We translated the English version of the Inventory of Individual Differences in the Lexicon (IIDL) into Chinese, and within a single survey session participants either completed the instrument either in both languages (N=151 bilingual participants) or twice in Chinese (N=94) or in English (N=82). Finally, additional bilingual participants (N=46) rated the perceived translation quality of each item. Variation in the cross-language correlations across items predicted perceived translation quality, however adjusting for same-language retest correlations resulted in significantly stronger indicators of perceived translation quality. The present study thus indicates the validity of McCrae et al.'s (1998) general method, and demonstrates that it can be extended to designs where all participants complete a single test session and can be applied to evaluate the quality of translations of single items.

Keywords: scale translation, translation quality, scale reliability, bilingual, within-session retest

Adjusting Bilingual Ratings by Retest Reliability Improves Estimation of Translation Quality

Evaluating the quality of scale translations is an important yet challenging task in cross-cultural psychology. Effective evaluation of translation quality is substantially handicapped by the intertwining influences of language differences, cultural differences, sample differences, or combinations of the three (Hulin, 1987; John, Goldberg, & Angleitner, 1984). Bilingual individuals who are proficient in two languages are often used to disentangle the influences (Butcher, 2004; Mallinckrodt & Wang, 2004; Sireci, 2004; Sperber, Devellis, & Boehlecke, 1994). They are asked to respond to items twice in different languages within a single study, so that language differences may be the only factor that results in the differences between the two language versions. High correlations between bilinguals' scores on the original and translated forms of the measure indicating the meaning of the scale has been preserved (Butcher, Mosch, Tsai, & Nezami, 2006; Costa, McCrae, & Kay, 1995; John et al., 1984; McCrae et al., 1998; Piedmont & Chae, 1997).

Although low correlations can indicate that the translated scale has not preserved the meaning of the original scale, they can also reflect unreliable (inconsistent) responses to the scale. Specifically, it is possible that correlations on two forms of the test will be low *even if the second form is a direct repetition of the first.* Therefore, to better estimate the cross-language equivalence of the original and translated items, the correlation of scores provided by bilingual participants across languages can be adjusted by their retest reliability when the scales are rated twice in the same language over the same measurement interval, as in the equation below:

Equation 1. Adjusted cross-language correlation: $\hat{\rho}_{\mathbb{X}_A\mathbb{X}_B(m)} = \frac{r_{X_A X_B(m)}}{\sqrt{r_{XX_A(m)} \times r_{XX_B(m)}}}$

Where X_A and X_B indicate observed scores on what we intend to interpret as "the same scale" X which has been translated into forms A and B (e.g., English and Chinese forms), *m*

indicates the *measurement interval* separating measurements of the scores being correlated. For instance, $r_{XX_{Eng}(3month)}$ indicates the retest correlation of scores on the English version of the scale over a three-month interval. Finally, $\hat{\rho}_{\mathbb{X}_A\mathbb{X}_{B(m)}}$ indicates the estimated correlation between *expected scores* on forms A and B within interval *m* – analogous to the *true score correlation* in classical test theory, and roughly interpretable as the correlation between the average scores on X_A and X_B that participants would obtain if they completed the forms a very large (conceptually infinite) number of times within the measurement interval (Lazarsfeld, 1959; Lord & Novick, 1968).

There is increasing evidence that retest correlations are particularly valuable estimates for use in reliability adjustments, as is done in Equation 1. For instance, McCrae, Kurtz, Yamagata, and Terracciano (2011) found retest correlations to better track scale validity coefficients than internal consistency statistics such as alpha. As shown by de Vries, Realo, and Allik (2016) and Lowman, Wood, Armstrong, Harms, and Watson (2018), retest correlations also help to resolve the vexing problem of how to estimate the reliability of single items, as scales of any length can be retested. But perhaps even more importantly, these studies have found retest correlations to better track item-level validity coefficients (e.g., self-other agreement, long-term stability) than other coefficients which fall in the family of internal consistency coefficients (e.g., the squared communality of the item within a multi-item scale; Wanous & Hudy, 2001; Denissen, Geenen, Selfhout, & Van Aken, 2008). Given the increasing understanding that single items almost invariably contain meaningful variance that is lost when aggregating items into multi-item scales (Mõttus, Kandler, Bleidorn, Riemann, & McCrae, 2017), this is an important advance for determining how to appropriately deal with issues of measurement unreliability at this level of analysis. At a more conceptual level, retest correlations more directly operationalize the definition of a

reliability coefficient as "the correlation of a measure with itself" (John & Soto, 2007, p. 464; Guttman, 1945; Lumsden, 1978).

Finally, Equation 1 operationalizes an understanding that the correlations between measures should be adjusted for unreliability *using the retest correlations of the measures over the same measurement interval (m)*. For instance, if two tests are measured 2 weeks apart, then one should use the 2-week retest correlations for reliability adjustments; if measured 30 minutes apart, then one should use the 30-minute retest correlations, and so on. When the measurement interval *m* is equated across the three correlations used in Equation 1 in this manner, we can interpret Equation 1 as a *counterfactual ratio* which indexes how much smaller the cross-language correlation of the test is than the correlation we would have obtained by instead repeating the tests twice in the same languages over the same measurement interval.

The above method for estimating the cross-language equivalence of scores was first explored in a study by McCrae and colleagues (1998). They asked a group of English-Chinese bilingual students to respond twice to the NEO-PI-R, once in the original language and once in the translated language two weeks later ($r_{X_{Eng}X_{Chi}(2week)}$). They also asked other bilingual students to rate the inventory twice in the same language (English or Chinese) over the same two-week interval, in order to obtain the same-language retest reliabilities of the scale ($r_{XX_{Eng}(2week)}$ and $r_{XX_{Chi}(2week)}$). Results indicate that some relatively low cross-language retest correlations may be due to the retest unreliability of the scale rather than translation inequivalence, because the disattenuated correlations were high after adjusting for the simple retest unreliability. For instance, ratings of the NEO Tender-Mindedness scale in English and Chinese collected two weeks apart correlated only at a level of $r_{X_{Eng}X_{Chi}(2week)}$ = .51, but were estimated to correlate at a level of $\hat{\rho}_{\mathbb{X}_{Eng}\mathbb{X}_{Chi}(2week)}$ = 1.07 after adjusting for

unreliability. According to the authors, this indicates that the meaning of the scale had been well-preserved across the original and translated scales.

Despite the strong psychometric logic for this method, which can be understood as a cross-cultural analogue of standard techniques for adjusting for measurement unreliability (e.g., John & Benet-Martinez, 2000; Schmidt, Le, & Ilies, 2003; Spearman, 1904, 1910), there are a number of practical limitations to the above procedure for estimating the quality of scale translations. These may account for the fact that this method does not appear to have been employed since McCrae and colleagues' 1998 study. First, McCrae and colleagues (1998) study suggested that scale ratings be made in different sessions to increase measurement independence, however this comes at considerable costs to experimenters and participants, where it may be difficult to get participants to return to a second testing session. Furthermore, separating the repeated measurements into a different session (e.g., two weeks later) will decrease both the correlations between the original and translated form of the measure and each measure's retest correlation relative to shorter intervals, as longer time intervals will typically decrease inter-item and retest correlations (e.g., Fraley & Roberts, 2005). As discussed by Wood and colleagues (2018), this may not decrease the *expected validity* of this method of adjusting for measurement unreliability, but should result in Equation 1 producing more unstable estimates of the scale translation quality. This occurs because underestimates of the population retest correlations, which are expected to occur through sample fluctuations, will result in larger over-adjustments for score inconsistency. This is perhaps evidenced by the existence of several "out-of-bound" estimates reported in McCrae et al.'s (1998) original investigation (i.e., 6 of the 30 $\hat{\rho}_{\mathbb{X}_{Eng}\mathbb{X}_{Chi}(2week)}$ estimates exceeded 1.00).

To address the above limitations, we propose to substantially reduce the time interval separating the first and second administration of the instrument by administering the measure

twice within the same survey session. The repetition of the instrument thus is separated by a mere 10 minutes in which participants rate other measures. As argued by Lowman and colleagues (2018) and Wood and colleagues (2018), this method of estimating *within-session retest correlations* provides feasible reliability estimates for operationalizing Equation 1 because retesting even over an interval of 10 minutes (in which participants rate many other items) should be sufficient to largely eliminate participants' memory of the specific answers they have given previously. Other properties of modern online surveys – such as the ability to easily randomize the order of measures and items and to prevent the possibility of looking back to one's previous answers – should further increase the independence of within-session repeated measurements. More concretely, similar to demonstrations by McCrae and colleagues (2011) and de Vries, Realo, and Allik (2016), these authors demonstrated that same-session retest correlations outperform internal-consistency estimators of reliability (e.g., coefficient alpha) by better tracking between-scale variation in properties expected to be impacted by measurement unreliability, like self-other correlations and long-term stability (e.g., 1-year).

In the present study, we will also address an important limitation to the method of estimating translation quality proposed by McCrae and colleagues (1998), as represented in Equation 1. Despite its intuitive psychometric logic: *it has never actually been demonstrated to result in estimates which better track the quality of the scale translation*. There are reasons that these adjustments may not achieve this result: if estimates of the three correlations needed to estimate $\hat{\rho}_{\mathbb{X}_A\mathbb{X}_B(m)}$ are sufficiently small in magnitude, or are estimated in sufficiently small samples, taking the ratio of these three correlations may introduce more bias than they remove. Consequently, we conduct the first study to our awareness to evaluate whether adjusting raw-score cross-language correlations by reliability estimates actually results in better predictors of the perceived quality of the scale translation. This was done by

having an independent sample of bilingual participants evaluate the extent to which the original and translated items are equivalent in meaning. Demonstrating this final point will serve as a crucial piece of evidence for establishing whether the reliability-adjustment represented in Equation 1 results in improved estimates of translation quality. If so, the approach should be more widely considered in cross-cultural methodology.

Method

Measurement Translation

The Inventory of Individual Differences in the Lexicon (IIDL; Wood, Nye, & Saucier, 2010) is an instrument designed to survey a wide range of individual differences where conceptually distinct traits are assessed by one item each (e.g., Block, 1961; Funder, Furr, & Colvin, 2000). It contains 84 items consisting of pairs of synonymous adjectives, such as "sociable, outgoing" and "smart, intelligent" on a scale with anchors ranging from 1 (*Extremely Uncharacteristic*) to 7 (*Extremely Characteristic*). The large number of items and broad range of item content make this inventory appropriate for our study.

We translated the IIDL into Chinese via the following steps. First, five research assistants from China who were fluent in English independently translated the English version into Chinese. Then, they met with two authors who are native Chinese speakers to finalize the Chinese translation. A back translation was conducted by a professional translator who is a native Chinese speaker majoring in English. Then, three authors (one native English speaker and two native Chinese speakers) met and modified the Chinese items based on the back-translation results.

English-Chinese Within-Session Retest

A total of 151 students from a large university in Singapore (84 females, 67 males; M[SD] age = 22.2[1.5] years) participated in our study for course credits. They reported being fluent in both English and Chinese when asked about their language fluency. Aside

from this there was Our study was conducted online and participants could not look back at their answers. All participants completed both the English and Chinese version of IIDL in a counterbalanced order. Between the two versions, 111 students completed 49 items related to life satisfaction (i.e., Satisfaction With Life Scale; SWLS) and another personality measure (i.e., Big-Five Inventory; BFI-44), and 40 students completed 198 items related to cultural beliefs, food preferences, and other personal characteristics (i.e., BFI-44).

English-English Within-Session Retest

Eighty-two students from a large university in Singapore (63 females, 18 males, 1 missing; M[SD] age = 20.6[1.6] years) participated in our study for course credits[1]. They reported being fluent in English when asked about their language fluency. Our study was conducted online and participants could not look back at their answers. Participants rated the English version of the IIDL twice. In between, participants rated approximately 110 items related to emotion (e.g.,, the Positive and Negative Affect Schedule; PANAS), well-being (SWLS), and other characteristics (BFI-44) before re-ratings the IIDL items.

Chinese-Chinese Within-Session Retest

Ninety-four students from a large university in Singapore (65 females, 29 males; M[SD] age = 18.8[1.6] years) participated in our study for course credits. They reported being fluent in Chinese when asked about their language fluency. Our study was conducted online and participants could not look back at their answers. Participants completed the Chinese version of the IIDL twice within a single testing session. In between, participants rated approximately 110 other items in Chinese related to emotion (e.g., PANAS), well-being (e.g., SWLS), and other personal characteristic (e.g., BFI-44).

Perceived Translation Quality

Finally, a group of 46 students from a large university in Singapore (35 females, 11 males; M[SD] age = 21.39[1.5] years) who reported being fluent in both English and Chinese

participated in the study for course credit. IIDL items were presented in both English and Chinese side-by-side in an online survey. Participants were asked to indicate how similar the English and Chinese items are in describing people or actions on a scale from 1 (*Not at all similar*) to 5 (*Essentially the same*). Higher *perceived translation quality* ratings thus indicate better preservation of the communicated meaning of the original item (Sperber et al., 1994). When considering the ratings from each of the 46 raters as 'indicators' of the perceived translation quality, the reliability of the similarity rating was high (α = 0.86). As an 'expected alternative form correlation', this indicates the expected correlation of these mean ratings with means obtained from a new set of 46 raters (Cronbach, 1951). This also indicates that the average inter-rater agreement regarding the ordering of 'translation quality' scores across the 84 IIDL items was .12.

Results

As shown in Table 1, the average English-Chinese within-session correlations was M(SD) =.55(.15), with the cross-language correlations ranging from a low of $r_{X_{Eng}X_{Chi}(d)}$ = .21 for the pair "pleasant, agreeable" / "和气的、随和的" to a high of .85 for the pair "short, little"/ "矮的、小个的." In addition, the average perceived translation quality was also high; M(SD)= 4.10(.27), indicating that bilingual participants perceived the English and Chinese versions of the IIDL items as generally having very similar meanings. Table 1 also indicates, consistent with McCrae and colleagues (1998) investigation, that adjusting cross-language correlations by same-language retest-correlations resulted in a small number of 'out-of-bound' correlations (i.e., $\hat{\rho}_{\mathbb{X}_A\mathbb{X}_B(d)}$ estimates exceeding 1).

As shown in Table 2, higher English-Chinese within-session correlations were associated with higher estimates of perceived translation quality; q = .35 (p < .01) [2]. Most importantly, when Chinese-Chinese and English-English within-session reliability was used

to adjust for unreliability in the English-Chinese within-session correlations, the correlation between the adjusted estimates and the perceived translation quality estimates increases to q = .47 (p < .01)[3]. Given the very high correlation between the rank-ordering of the raw and adjusted English-Chinese retest correlations across items, q = .90, this was a statistically significant difference in the relative validity of the two estimates as indicators of the perceived translation quality by Steiger's (1980) test of differences in dependent correlations (Z = 2.69, N = 84, p < .01).[4] This result indicates that adjusting the raw correlation between English-Chinese scores by their retest-reliabilities (administered twice within the same language) does in fact result in a better indicator of the equivalence of items across languages.

A graphical representation of these results is given in Figure 1. As this figure illustrates, despite the high q = .90 correlation between the overall rank-ordering of items before and after adjusting for retest consistency, the rank-ordering of items estimated as having the highest correlations before and after this adjustment changed considerably.

Discussion

The current study presents a critical evaluation and extension of a method that has been used to evaluate the quality of item translations in cross-cultural research. Researchers have shown that when bilinguals completed the same measure in different languages, the raw-score cross-language correlation can be divided by the same-language retest correlation over the same interval to estimate the quality of the translation (McCrae et al, 1998). Perhaps the most important contribution of the current research is to provide the first empirical evidence that the estimated correlations produced by this means of adjusting for score unreliability do in fact result in better indicators of translation quality, by showing that they outperform raw-score correlations in predicting the extent to which items are *perceived* as similar in meaning by bilingual participants. Our results indicate that cross-language within-session retest

correlations can provide accurate estimates of translation quality, and that the level of prediction may be improved by using the adjustment for unreliability formula given in Equation 1.

The approach used in the present research also helps to show how the technique developed by McCrae and colleagues (1998) can be more practically implemented in several ways. First, we demonstrated that this method can be used over shorter intervals – in particular: when individuals have completed the inventory twice (in the same or different languages) within a single survey session. Past applications of this general approach indicated that participants should complete two or more sessions separated by a relatively long period, such as the two-week interval used by McCrae and colleagues. The ability to collect all necessary data from participants who have only completed a single survey session reduces the experimenter and participant resources necessary to complete the study, which should make it easier to obtain larger sample sizes. Additionally, as correlations tend to decrease in magnitude as the scores being correlated are separated farther in time, collecting the measures necessary to adjust for score unreliability within a single session has the expected effect of increasing inter-item correlations (Lowman et al., 2018). Both of these features will serve to result in more stable estimates of the translation quality.

Further, an intriguing feature of the adjustment for unreliability given in Equation 1 is that it can be applied to measures of any length. This means that it can be used to evaluate not just the translation quality of multi-item scales, but also the translation quality of every item within the scale. The results of the present analysis shown in Table 2 and Figure 1 support the broader argument that adjusting for measurement unreliability using this method results in significantly improved estimates of the quality of translations at the level of single items. This is important as it affords the opportunity of evaluating which particular items in a broader multi-item scale may be responsible for different performance of translated forms.

Limitations and Future Directions

The research design used to evaluate the equivalence of scales across translation relaxed certain study design features used by McCrae and colleagues (1988). Specifically, they limited their entire analysis to bilingual participants, who were randomly assigned to complete the survey either in Chinese or English during the first administration, and then randomly assigned to do so again during the second administration. In contrast, the "same-language retest correlations" used here, and reported in Table 1, were collected from participants that were not necessarily bilingual (i.e., those in the English-English group). As bilinguals may be very different from the monolingual groups, in the current study, the same-language retest correlations will be influenced by sources of error due to particularities of the monolingual group in addition to errors due to time sampling, while the cross-language correlations will be influenced not just to content sampling (English vs. Chinese) but also particularities of the bilingual group. In contrast, McCrae et al.'s approach presumably reduces confounding sources of variance that could affect the validity of disattenuated estimates. However, despite these potential limitations, we observed that adjusting cross-language correlations using these retest reliabilities estimated from the single language groups nonetheless improved the quality of translations perceived by an independent bilingual sample.

Some adjusted correlations exceeded 1.00. We believe such observations can mostly be attributed to the modest sample sizes used to estimate some of the components of Equation 1. Specifically, the components in the denominator of Equation 1 were estimated using sample sizes near N = 100, which can cause estimates to fluctuate considerably. For instance, the Chinese translation of the English IIDL item "tired, exhausted", "疲劳的、精疲力尽的" showed the *lowest* Chinese-language retest correlation across all items ($r_{XX_{Chi}(d)}$ = .39), which in turn resulted in the *highest* disattenuated estimate of translation quality ($\hat{\rho}_{\mathbb{X}_{Eng}\mathbb{X}_{Chi}(d)}$

= 1.20). Several of the other items that were estimated to have adjusted translation-quality estimates exceeding $\hat{\rho}_{\mathbb{X}_A\mathbb{X}_B(d)}$ = 1 also showed at least one same-language reliability below a .60 magnitude. Although adjusted correlations exceeding $\hat{\rho}_{\mathbb{X}_A\mathbb{X}_B(d)}$ = 1 are expected to occur regularly when two forms of a test are *perfectly* parallel (i.e., the translated scale provides *exactly* the same ordering of expected scores as the original scale; Charles, 2005), this condition should be rarely met, and so 'out-of-bound' estimates should become infrequent as sample sizes increase.

Despite these limitations, we nonetheless found that adjusting the bilingual cross-language correlations by the same-language retest correlations in each language resulted in improved estimates of translation quality, as judged by an independent sample of bilingual participants. Furthermore, we actually observed *fewer* adjusted correlations greater than 1.00 within the present method than reported in McCrae and colleagues (1998) – i.e., only 6 of 84 items, or 7%, compared to 6 of 30, or 20%. This indicates that the method may be relatively robust across the condition of whether the cross-language correlations and same-language retest-correlations are all collected with bilingual participants and estimated at the scale or item level. Some of the other features of the present study – especially the shorter interval separating retests – help to compensate for such limitations. This is useful given the fact that large samples of bilingual participants may be difficult to recruit for such scale evaluation studies.

Even if the reliability-adjusted translation quality of an item is *truly* and *appropriately* estimated near unity ($\rho_{\mathbb{X}_A\mathbb{X}_B(d)} \approx 1$), items with low within-session retest correlations in a given language (e.g., $\rho_{XX_A(d)} \approx .50$) may still be considered problematic. We can interpret this situation as meaning that we would obtain the same ordering of *expected scores* on the measures – i.e., the scores participants would receive on the test in each language if averaging their responses across a large (conceptually infinite) number of repeated

assessments – even though they do not provide a consistent ordering of scores across *single assessments*. Although within-session retest correlations may serve as particularly useful reliability estimates of psychological states (Lowman et al., 2018), it seems likely that low within-session retest correlations may often indicate that participants have interpreted a specific item with reference to their current state, which may fluctuate considerably even within a 15-minute retest interval. For instance, the .39 within-session retest correlation for the item 'tired, exhausted' may come from participants interpreting the item as a state (how tired I am *right now*) rather than as a trait (how tired I tend to be *generally*). The low correlation could reflect meaningful fluctuation in state-level fatigue during the course of completing the survey. If the goal is for participants to provide *trait* ratings, low within-session retest correlations may help to identify items that are not interpreted in the desired manner. On the other hand, many traits that are considered an important aspect of personality may pertain to content that participants simply are unable to report consistently, perhaps due to the breadth or more abstract (e.g., less observable) nature of the trait. For instance, the items "afraid, scared" "害怕的、怕的" and "kind-hearted, caring" "好心的、关怀的" showed modest within-session reliabilities in both languages, but previous studies have also indicated that participants may simply respond to items related to the Big Five domains of neuroticism and agreeableness more inconsistently (Gnambs, 2014; Wood & Wortman, 2012).

Conclusion

The current study helps to better establish the value of a technique for evaluating the quality of translated items first developed by McCrae and colleagues (1998), and which can be understood as a cross-cultural application of a more general method for evaluating test equivalence (e.g., Spearman, 1904; Lord & Novick, 1968; John & Benet-Martinez, 2000). To our awareness, our results provide the first empirical evidence that adjusting observed cross-

language correlations by their estimated retest reliability over the same measurement interval results in a significantly strong indicator of the quality of the scale translation.

The results further show that these reliability-adjusted estimates of translation quality can be estimated for scales of any length – including single items – and can be validly estimated from repetitions of the test within a single larger survey session. Both of these features serve to increase the practicality and utility of this method for cross-cultural researchers.

Footnotes:

1. Participants in the Chinese-Chinese (English-English) groups were not asked how fluent they were in English (Chinese). However, because the language of instruction at all Singaporean universities is English, all students must be proficient in English to be admitted. Thus, participants in the Chinese-Chinese group were effectively bilingual. The majority of participants in the English-English group were also likely to be bilingual given that an estimated 73.6% of students at the university are bilingual (Singapore Department of Statistics, 2011).

   There were other estimates of English-English within-session retest correlations from different samples in the U.S. (Wood, et al., 2017). Across items, the estimates from the Singapore sample correlated with those of three other samples in the magnitude of *q* = .46, .54, and .60, *p*s<.05. In addition, the M(SD) within-session retest-correlation estimates, $r_{XX_{Eng}(d)}$, across all items for the SG sample is .78(.09), while that of the other three samples is .68(.10), .72(.12), and .76(.10), respectively. This suggests that the results from our sample are comparable to those from other studies.
2. Following conventions developed by Cattell (1952), within this paper we use *q* to indicate correlations at the 'between-test' level of analysis (e.g., between scale or item properties) and reserve *r* to indicate correlations at the 'between-person' level of analysis.
3. If the six items with $\hat{\rho}_{\mathbb{X}_{Eng}\mathbb{X}_{Chi}(d)}$ values exceeding 1 were rescored as having values of 1, this correlation increased very slightly to .48.
4. It is important to note that this is not a statistically necessary result. Specifically, it is true that adjusting for unreliability will result in expected score (or true score) correlations that must necessarily be larger in magnitude than raw-score correlations – i.e., $\hat{\rho}_{\mathbb{X}\mathbb{Y}} > r_{XY}$ for any and all test pairs that have less than perfect reliabilities. However, here we are discussing how these correlational indices of cross-language score consistency in turn correlate with other measurement properties at the between-stimulus or between-item level of analysis – in this case, the perceived translation quality of the items. If the reliability estimates used to adjust

for raw-score correlations are invalid (for instance, if they represent random variables), then $\hat{\rho}_{\mathbb{X}_A\mathbb{X}_B(d)}$ estimates could show *significantly lower* correlations with perceived translation quality across items by being infused with more invalid variance than simple $r_{X_AX_B(d)}$ raw-score correlations.

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research* (Vol. 457). Springfield, IL: Thomas.

Butcher, J. N. (2004). Personality assessment without borders: Adaptation of the MMPI-2 across cultures. *Journal of Personality Assessment*, *83*, 90-104.

Butcher, J. N., Mosch, S. C., Tsai, J., & Nezami, E. (2006). *Cross-cultural applications of the MMPI-2*. Washington, DC: American Psychological Association.

Cattell, R. B. (1952). The three basic factor-analytic research designs—their interrelations and derivatives. *Psychological Bulletin*, *49*, 499-520.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*, 206-226.

Costa Jr, P. T., McCrae, R. R., & Kay, G. G. (1995). Persons, places, and personality: Career assessment using the Revised NEO Personality Inventory. *Journal of Career Assessment*, *3*, 123-139.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

de Vries, R. E., Realo, A., & Allik, J. (2016). Using personality item characteristics to predict single-item internal reliability, retest reliability, and self–other agreement. *European Journal of Personality*, *30*, 618–636.

Denissen, J. J., Geenen, R., Selfhout, M., & Van Aken, M. A. (2008). Single-item Big Five ratings in a social network design. *European Journal of Personality*, *22*, 37–54.

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*, 60-74.

Funder, D. C., Furr, R. M., & Colvin, C. R. (2000). The Riverside Behavioral Q-sort: A tool for the description of social behavior. *Journal of personality*, *68*, 451-489.

Gnambs, T. (2015). Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences, 84*, 84-89.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255–282.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of Cross-Cultural Psychology*, *18*, 115-142.

John, O. P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339–369). New York: Cambridge University Press.

John, O. P., Goldberg, L. R., & Angleitner, A. (1984). Better than the alphabet: Taxonomies of personality-descriptive terms in English, Dutch, and German. In H. Bonarius (Ed.), *Personality psychology in Europe* (Vol. vol. 1: Theoretical and empirical developments, pp. 83-100). Lisse: Swets & Zeitlinger.

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, R. F. Krueger, R. W. (Eds.), *Handbook of research methods in personality psychology* (pp. 461–494). New York: Guilford.

Lazarsfeld P.F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A Study of Science* (pp. 476–543). New York: McGraw-Hill.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lowman, G. H., Wood, D., Armstrong, B. F. I., Harms, P. D., & Watson, D. (2018). Estimating the reliability of motion measures over very short intervals: The utility of within-session retest correlations. *Emotion*. doi:10.1037/emo0000370

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31*, 19–26.

Mallinckrodt, B., & Wang, C. C. (2004). Quantitative methods for verifying semantic equivalence of translated research instruments: A Chinese version of the Experiences in Close Relationships Scale. *Journal of Counseling Psychology, 51*, 368-379.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*, 28–50.

McCrae, R. R., Yik, M. S., Trapnell, P. D., Bond, M. H., & Paulhus, D. L. (1998). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology*, *74*, 1041-1055.

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*, 474–490.

Piedmont, R. L., & Chae, J. H. (1997). Cross-cultural generalizability of the five-factor model of personality: Development and validation of the NEO PI-R for Koreans. *Journal of Cross-Cultural Psychology*, *28*, 131-155.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*, 206–224.

Singapore Department of Statistics. (2011). *Singapore census of population 2010, statistical release 1: Demographic Characteristics, education, language and religion.* Retrived

from https://www.singstat.gov.sg/-/media/files/publications/cop2010/census_2010_release1/cop2010sr1.pdf

Sireci, S. G. (2004). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K., Hambleton, P. F., Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp.117-138). Oxford, UK: Psychology Press.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology, 15*, 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-cultural translation methodology and validation. *Journal of Cross-Cultural Psychology*, *25*, 501-524.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251.

Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods, 4*, 361–375.

Wood, D., Harms, P. D., Lowman, G., Soto, C. J., Qiu, L., John, O. P., & Lu, J. (2018). *Evaluating the utility of within-session retest correlations as reliability estimates*. Manuscript submitted for publication, University of Alabama, Tuscaloosa, AL.

Wood, D., Nye, C. D., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive trait markers from the English lexicon. *Journal of Research in Personality*, *44*, 258–272.

Wood, D., & Wortman, J. (2012). Trait means and desirabilities as artifactual and real sources of differential stability of personality traits. *Journal of Personality, 80*, 665-701.

Table 1. Item-level estimates of same-language and cross-language correlations and perceived translation quality.

			Within-session retest correlations					
#	Original English item	Chinese Translation	Chinese-Chinese $r_{XX_{Chi}(d)}$ (N=94)	English-English $r_{XX_{Eng}(d)}$ (N=82)	English-Chinese $r_{X_{Eng}X_{Chi}(d)}$ (N=151)	Reliability-adjusted English-Chinese, $\hat{\rho}_{\mathbb{X}_{Eng}\mathbb{X}_{Chi}(d)}$	Perceived translation quality (N=46) M	SD
1	afraid,scared	害怕的、怕的	.60	.69	.77	1.20	4.33	0.79
2	tired,exhausted	疲劳的、精疲力尽的	.39	.80	.67	1.20	4.17	1.02
3	bashful,shy	腼腆的、害羞的	.46	.79	.70	1.16	4.09	0.97
4	kind-hearted,caring	好心的、关怀的	.56	.63	.65	1.09	4.33	0.87
5	smart,intelligent	聪明的、智慧的	.70	.79	.76	1.02	4.35	0.85
6	direct,straight-forward	直接的、直截了当的	.63	.76	.71	1.02	4.15	0.79
7	well,healthy	良好的、健康的	.55	.82	.67	.996	4.37	0.74
8	lonely,lonesome	孤独的、孤寂的	.63	.80	.69	.98	4.22	0.87
9	short,little	矮的、小个的	.87	.87	.85	.98	4.15	0.94
10	sad,unhappy	悲伤的、不开心的	.72	.75	.71	.97	4.13	0.83
11	feminine,unmasculine	女性化的、不阳刚的	.79	.90	.80	.95	3.59	1.05
12	good-looking,attractive	好看的、吸引人的	.77	.95	.81	.95	3.98	0.91
13	prompt,punctual	快捷的、守时的	.67	.86	.71	.94	4.11	0.71
14	excited,enthusiastic	兴奋的、热情的	.63	.74	.63	.93	4.18	0.78
15	loud,noisy	大声的、吵闹的	.79	.88	.78	.93	4.41	0.78
16	funny,amusing	好笑的、滑稽的	.78	.81	.73	.92	3.76	0.80
17	likeable,well-liked	讨喜的、受欢迎的	.70	.79	.68	.92	4.15	0.79
18	lively,playful	活泼的、调皮的	.66	.75	.65	.92	4.41	0.65
19	brave,adventurous	勇敢的、爱冒险的	.82	.79	.73	.91	4.33	0.90
20	unfriendly,cold	不友善的、冷淡的	.61	.76	.62	.91	4.33	0.83
21	weird,strange	古怪的、奇怪的	.82	.82	.75	.91	4.22	0.84
22	independent,self-sufficient	独立的、自给自足的	.61	.81	.64	.91	4.28	0.91
23	giving,generous	大方的、慷慨的	.69	.55	.55	.90	4.37	0.83
24	wealthy,well-to-do	富裕的、富有的	.87	.79	.75	.90	4.20	0.69
25	sociable,outgoing	社交的、外向的	.85	.87	.77	.89	4.04	0.87
26	positive,optimistic	积极的、乐观的	.79	.79	.70	.88	4.02	0.98
27	competent,capable	能干的、有能力的	.66	.79	.63	.87	4.36	0.72
28	lucky,fortunate	好运的、幸运的	.78	.80	.69	.87	4.28	0.69
29	beautiful,pretty	美丽的、漂亮的	.85	.91	.76	.87	4.35	0.85
30	happy,joyful	开心的、喜悦的	.67	.82	.64	.86	4.46	0.75
31	dominant,controlling	强势的、控制的	.66	.84	.63	.85	4.09	0.86
32	determined,persistent	有决心的、坚持的	.63	.67	.55	.85	4.37	0.93
33	dumb,stupid	笨的、愚蠢的	.62	.74	.58	.85	4.30	0.73
34	cheap,stingy	抠门的、吝啬的	.71	.72	.60	.84	3.67	0.92

35	disorganized,messy	混乱的、凌乱的	.60	.82	.58	.83	3.96	0.87
36	jealous,possessive	嫉妒的、占有欲强的	.74	.83	.65	.83	3.93	0.93
37	sarcastic,critical	讥讽的、批评的	.69	.85	.63	.82	3.89	0.99
38	truthful,honest	真实的、诚实的	.58	.62	.48	.80	4.33	0.82
39	confident,self-assured	自信的、自我肯定的	.78	.83	.64	.80	4.28	0.83
40	youthful,young	青春的、年轻的	.78	.75	.61	.80	4.30	0.79
41	polite,courteous	礼貌的、有礼的	.56	.66	.48	.79	4.35	0.74
42	selfish,self-centered	自私的、自我为中心的	.67	.82	.59	.79	4.59	0.69

			Within-session retest correlations					
#	Original English item	Chinese Translation	Chinese-Chinese $r_{XX_{Chi}(d)}$ (N=94)	English-English $r_{XX_{Eng}(d)}$ (N=82)	English-Chinese $r_{X_{Eng}X_{Chi}(d)}$ (N=151)	Reliability-adjusted English-Chinese, $\hat{\rho}_{\mathbb{X}_{Eng}\mathbb{X}_{Chi}(d)}$	Perceived translation quality (N=46) M	SD
43	creative,imaginative	有创造力的、有想象力的	.81	.85	.65	.79	4.43	0.75
44	egotistical,conceited	自大的、自负的	.67	.63	.51	.78	3.96	0.99
45	influential,prominent	有影响力的、显赫的	.68	.73	.55	.78	4.11	0.90
46	ordinary,average	平常的、一般的	.51	.74	.48	.78	4.20	0.98
47	slim,slender	苗条的、修长的	.84	.91	.68	.78	4.17	0.82
48	hot-tempered,short-tempered	暴躁的、易怒的	.69	.88	.61	.78	4.09	0.78
49	conservative,traditional	保守的、传统的	.71	.85	.60	.77	4.45	0.70
50	inconsiderate,rude	不考虑他人的、无礼的	.69	.61	.49	.76	4.09	0.94
51	cruel,abusive	残忍的、虐待的	.65	.49	.43	.76	4.15	0.97
52	thankful,grateful	感谢的、感恩的	.62	.75	.51	.75	4.48	0.72
53	radical,rebellious	激进的、叛逆的	.74	.80	.57	.74	3.89	1.04
54	tense,anxious	紧张的、焦虑的	.64	.76	.49	.71	4.13	0.83
55	ashamed,humiliated	惭愧的、感到羞辱的	.58	.72	.46	.71	4.13	0.83
56	relaxed,calm	放松的、平静的	.63	.79	.49	.70	4.30	0.66
57	admirable,impressive	令人钦佩的、令人印象深刻的	.77	.75	.53	.70	3.91	1.09
58	assertive,bold	断言的、大胆的	.68	.75	.49	.69	4.11	0.92
59	affectionate,loving	情深的、有爱的	.64	.85	.50	.68	3.72	1.05
60	dependable,reliable	可靠的、可信赖的	.67	.49	.39	.68	4.54	0.62
61	efficient,thorough	高效的、彻底的	.67	.69	.46	.68	3.98	0.95
62	awkward,clumsy	笨拙的、不灵活的	.69	.85	.51	.67	3.39	1.16
63	great,terrific	很好的、很棒的	.61	.70	.43	.66	4.26	0.71
64	hard-working,productive	勤奋的、高产的	.64	.76	.46	.66	3.76	1.02
65	impulsive,spontaneous	冲动的、即兴的	.71	.76	.48	.65	4.09	0.94
66	faithful,loyal	忠实的、忠诚的	.58	.83	.44	.64	4.39	0.80
67	stable,well-adjusted	稳定的、完全适应了的	.65	.66	.41	.62	4.13	0.91
68	close-minded,narrow-minded	思想封闭的、思维狭隘的	.69	.51	.35	.59	4.09	0.76

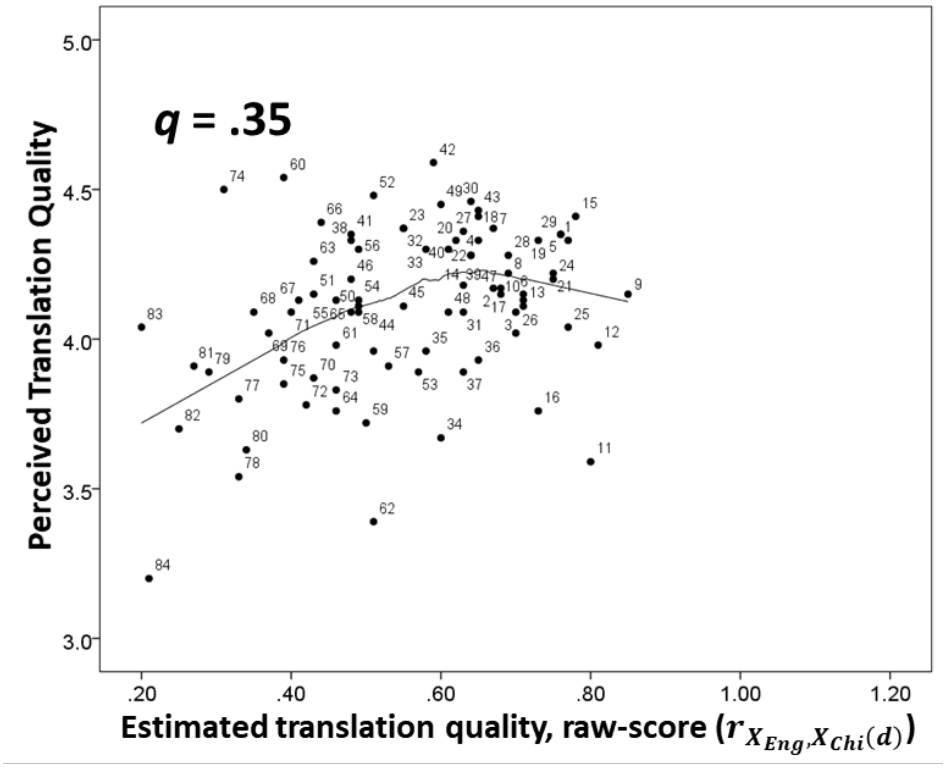
69	careful,cautious	仔细的、谨慎的	.66	.62	.37	.58	4.02	1.04
70	good-for-nothing,insane	一无是处的、发疯的	.72	.76	.43	.58	3.87	1.07
71	strict,firm	严格的、坚定的	.69	.74	.40	.56	4.09	0.84
72	exciting,fascinating	令人兴奋的、迷人的	.71	.86	.42	.54	3.78	0.96
73	retarded,senile	痴呆的、智力衰退的	.86	.85	.46	.54	3.83	1.16
74	undependable,unreliable	不可靠的、不可信赖的	.62	.58	.31	.52	4.50	0.66
75	skilled,skillful	技能熟练的、技艺精湛的	.69	.80	.39	.52	3.85	0.92
76	trusting,unsuspicious	相信人的、不多疑的	.79	.72	.39	.52	3.93	0.88
77	practical,sensible	实际的、合理的	.59	.71	.33	.51	3.80	1.02
78	angry,hostile	生气的、有敌意的	.70	.71	.33	.47	3.54	1.19
79	casual,informal	随意的、不正式的	.59	.68	.29	.46	3.89	0.95
80	temperamental,touchy	易怒的、过分敏感的	.72	.85	.34	.44	3.63	1.04
81	evil,corrupt	邪恶的、腐败的	.67	.71	.27	.39	3.91	1.03
82	crabby,grouchy	脾气坏的、有气的	.73	.70	.25	.35	3.70	0.84
83	pleasant,agreeable	和气的、随和的	.72	.56	.20	.31	4.04	0.87
84	hard,rough	坚硬的、铁石心肠的	.69	.69	.21	.31	3.20	1.05

Note. Items are ordered by the disattenuated English-Chinese within-session retest correlations. The reliability-adjusted value of English-Chinese within-session retest correlations was calculated by dividing the raw English-Chinese within-session retest correlations by the square-root of the product of the Chinese-Chinese and English-English within-session retest correlations.

Table 2. Correlations between within-session retest correlations and perceived translation quality

Item property	$r_{XX_{Chi}(d)}$	$r_{XX_{Eng}(d)}$	$r_{X_{Eng}X_{Chi}(d)}$	$\hat{\rho}_{\mathbb{X}_{Eng}\mathbb{X}_{Chi}(d)}$
Within-Session Retest Correlations				
Chinese-Chinese ($r_{XX_{Chi}(d)}$)	--			
English-English ($r_{XX_{Eng}(d)}$)	.34**	--		
English-Chinese ($r_{X_{Eng}X_{Chi}(d)}$)	.27*	.58**	--	
Adjusted English-Chinese ($\hat{\rho}_{\mathbb{X}_{Eng}\mathbb{X}_{Chi}(d)}$)	-.12	.30**	.90**	--
Perceived Translation Quality	-.16	-.11	.35**	.47**

Note. * $p < .05$, ** $p < .01$. The scores being correlated are given in Table 1; column labels are given in the corresponding rows. The disattenuated English-Chinese within-session retest correlations was calculated by dividing the raw English-Chinese within-session retest correlations by the square-root of the product of the Chinese-Chinese and English-English within-session retest correlations (Equation 1).



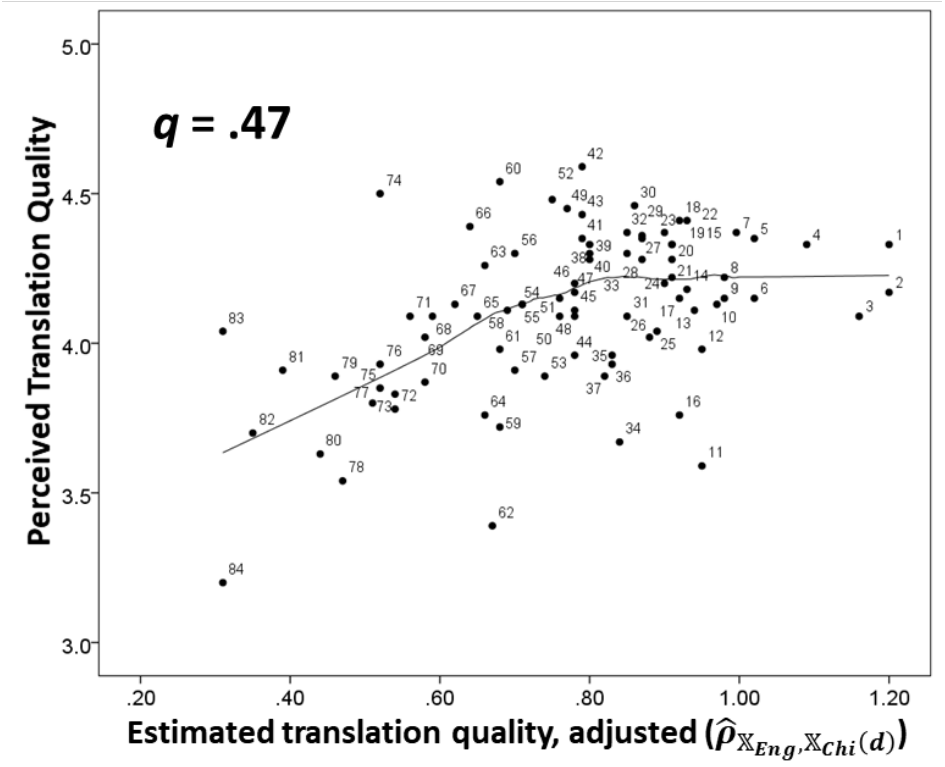


Figure 1. Relationships between within-session raw-score estimated translation quality ($r_{X_{Eng},X_{Chi}(d)}$) and reliability-adjusted translation quality ($\hat{\rho}_{\mathbb{X}_{Eng},\mathbb{X}_{Chi}(d)}$) with perceived translation quality. Note that items are labelled within the scatter plot by their row number in Table 1.