ELIA
Estudios de Lingüística Inglesa Aplicada

KEY CONCEPTS IN APPLIED LINGUISTICS/CONCEPTOS CLAVE DE LA LINGÜÍSTICA APLICADA
Reliability: What do We Mean When We (don't) Talk About It?

Shujing Zhao
Northern Arizona University
sz425@nau.edu

Luke Plonsky
Northern Arizona University
luke.plonsky@nau.edu

1. Introduction

Imagine we plan to investigate students' second language (L2) learning motivation in our classroom. One of the first decisions that we need to make is what tool to use for measuring L2 motivation, which usually leads to the next question: How do we know the tool is sufficiently reliable so that the measurement results can be trusted?

Just like motivation, most constructs of interest in applied linguistics cannot be observed directly. Therefore, tools and instruments are designed by researchers with an aim of measuring such so-called 'latent constructs' indirectly, and this is where reliability comes into play (McNeish, 2018; Plonsky & Derrick, 2016). Reliability

refers to the consistency that the items within the instrument show when measuring a given construct, providing evidence that the responses consistently and stably reflect the same thing (not necessarily the focal construct though, which would be part of a construct validity argument). Reliability is not an inherent, absolute trait of an instrument; instead, the reliability estimate is sourced from a specific set of responses and may change when the same instrument is used among different populations under different contexts.

We can see the importance of reliability from two perspectives. Firstly, inside the scope of a study, the instrument is expected to represent the "true value" of participants' performance. Let us take the example of L2 motivation again: If the instrument is reliable, the same group of students should receive similar scores when assessed multiple times under similar conditions. Otherwise, the scores can change drastically not because motivation has shifted but because of the inconsistency shown in the instrument. In other words, we hope that the responses received are not due to random chance or measurement error (i.e., the part of results that does not reflect the target construct): "If it isn't worth mentioning well, then it isn't worth measuring at all" (Cortina et al., 2020, p. 2). Using unreliable instruments might also produce larger standard deviations and consequently smaller effect sizes (e.g., Cohen's *d*), as well as lower probabilities of statistical findings (Larson-Hall & Plonsky, 2015; Plonsky & Derrick, 2016). Likewise, the accuracy of correlations between variables can be further negatively impacted (i.e., attenuated) by the low reliability of instruments used, meaning that the observed relationship appears weaker than it truly is. In such cases, researchers may misinterpret the results as indicating a smaller effect rather than attributing it low reliability. Secondly, beyond the scope of an individual study, reliability estimates offer a valuable point of reference to future researchers who might adopt the same instrument and who might benefit from guidance for interpreting their own data. Meta-analysts also benefit from reported reliability estimates which can be used to correct for attenuation in

primary studies to calculate a more precise estimate of the population effects of interest.

Despite the necessity of understanding, reporting, and interpreting reliability, a number of syntheses and meta-analyses have shown that applied linguists tend to under-report reliability estimates and/or to over-rely on reliability indices that are familiar but that might not be best suited to the data at hand (Al-Hoorie & Vitta, 2019; Plonsky & Derrick, 2016; Sudina, 2023). Plonsky and Derrick (2016) examined 14 syntheses of different L2 subdomains, and the percentage of studies reporting reliability estimates varies from 6% to 64%, which is far from ideal. Besides, interpretation of the reported reliability estimate is also usually omitted, "as if it were an item to tick off a list of submission guidelines rather than a meaningful source of information and interpretive value" (Larson-Hall & Plonsky, 2015, p. 141). In fact, taking a closer look at statistical literacy, especially knowledge about reliability, among applied linguists, we notice that we might not know as much as expected. Statistical knowledge was investigated in Loewen et al. (2014), and the factor analysis showed that reliability had the highest loading on a factor identified as "advanced statistics knowledge", which encompasses more sophisticated techniques such as Rasch analysis and structural equation modeling. Similarly, although applied linguistics students report relatively high ability of interpreting reliability, their self-rated ability to use this statistical concept tends to be lower, indicating that there is still a long way to go before the field fully understands what reliability estimates indicate and how to use and interpret them in a meaningful way (Gonulal et al., 2017; Zhang & Han, 2024). Moreover, the over reliance on one very familiar reliability index, Cronbach's alpha, is also problematic (see reasons in the following section), calling for more attention on the issue of reliability.

In the following sections of the article, therefore, we try to help deal with the issue by explaining why alpha is not an "one-size-fits-all" index, why omega can be a stronger alternative, how correlation can be attenuated due to reliability issues and how attenuation can

be corrected, and how reliability estimates from individual studies are synthesized in reliability generalization meta-analysis (RGM) to present a big picture of the field.

2. Cronbach's Alpha

Cronbach's alpha is the most commonly used measure of internal consistency reliability in behavioral sciences, including applied linguistics (Cortina et al., 2020; Larson-Hall & Plonsky, 2015; McNeish, 2018). However, like other statistical indices, alpha, as it's commonly known, carries certain assumptions which are rarely discussed and often (easily) violated. Specifically, to arrive an accurate estimate, alpha assumes (a) tau equivalence, (b) uncorrelated errors, (c) unidimensionality, and (d) normally distributed continuous data. We will go over the four assumptions very briefly and examine why they are often unmet.

1. Tau equivalence. When tau equivalence, or true-score equivalence, is satisfied, all items on a scale contribute equally to the underlying construct being measured and have identical factor loadings in a factor analysis. Tau equivalence, therefore, can be highly difficult to achieve as items in most instruments developed in applied linguistics capture and correlate with the measured construct to different extents.

2. Uncorrelated errors. Measurement error appears when there is discrepancy between the observed responses and the true value of results, which bring non-construct-relevant noise into observed values (McKay & Plonsky, 2021). To meet the assumption of uncorrelated errors, one needs to ensure that each item's measurements error is independent of the others. However, errors correlate when there is some systematic influence other than the measured construct itself affecting the items, such as item overlap, unclear wording, and even

the physical environment where the measurement takes place, requiring researchers to exercise caution when designing and implementing instruments (Cortina et al., 2020; McKay & Plonsky, 2021; McNeish, 2018).

3. Unidimensionality. As an index of internal consistency, alpha does not indicate unidimensionality, which represents homogeneity and assumes that all items on a certain scale measure the same underlying construct (Schmitt, 1996). Therefore, when one is not sure about the existence of any other construct or sub-construct within the scale, estimates of alpha might misestimate the scale's internal consistency.

4. Normally distributed continuous data. When items are presented in a Likert scale, which is commonly used in the field, the responses are discrete, and the covariances among items will be weakened if they are treated as continuous when computing alpha, leading to underestimation of the reliability estimate (McNeish, 2018).

From the explanation above, it is obvious that alpha might not be a good fit in many applied linguistics studies as the rigid assumptions constrain our ability to acquire a precise estimate of reliability (Cortina et al., 2020; Kelley & Pornprasertmanit, 2016; McNeish, 2018). In order to increase the precision of our research, we suggest that applied linguists make informed decisions on reliability estimation method based on study design and statistical knowledge instead of convenience, familiarity, and reflex. Meanwhile, we also provide below an alternative to Cronbach's alpha in the form of the omega coefficients.

3. Omega Coefficients

Compared with alpha, one of omega's advantages is that it relaxes the assumption of tau equivalence, meaning that it allows different

factor loadings of items and that items can show different sensitivities to the measured construct. Another advantage of omega is that it leaves room for multidimensionality and can be applied to scales that measure more than a single construct. In this section, we will introduce two omega coefficients, omega total and hierarchical omega[1], compare their application with examples, and explain how to interpret omega coefficients.

By definition, omega total estimates reliability through the variance "attributable to neither random error nor individual items" (Cortina et al., 2020, p. 20). In other words, omega total is calculated as the proportion of variance in observed results that can be explained by all common factors, including both general and group factors. Let us go back to the example of measuring L2 motivation to demonstrate this concept. If we simplify the construct of L2 motivation and assume it includes two sub-constructs in the scale, extrinsic motivation and intrinsic motivation, the general factor would be "L2 motivation", and the group factors would be "extrinsic motivation" and "intrinsic motivation". In this case, omega total provides an estimate of how reliable the scale is by taking all items into account, including both the overall L2 motivation and the two specific types of motivation (i.e., extrinsic and intrinsic motivation). One possible issue with omega total is that it considers the group factors, which might interfere with our focus on the general factor, usually the focal construct to be measured. For this reason, we can define reliability from another perspective, termed as hierarchical omega.

Hierarchical omega attempts to separate variance caused by group factors (i.e., subconstructs) and estimates reliability for a single general factor that dominates all items. Group factors are not neglected in hierarchical omega; instead, the probability of their

[1] The two omega coefficients discussed in this article both allow multidimensionality. See a unidimensional version of omega coefficient in Cho & Kim (2015). Unidimensional omega is also discussed in McNeish (2018) termed as "omega total", hence different from our definition here.

existence is acknowledged, and this omega coefficient aims to isolate group factors' influence from the general factor's influence, providing a more accurate measure for capturing the intended overarching construct (Kelley & Pornprasertmanit, 2016). Revisiting the example of an L2 motivation scale, we can see that hierarchical omega presents how reliably the scale measures just the overall motivation (i.e., general factor), without focusing on extrinsic and intrinsic motivation (i.e., group factors) specifically. Therefore, hierarchical omega is a better fit when a scale is designed to measure a single, overarching general factor and when the designer has not been sure about whether the scale is multidimensional or not, as this coefficient is an estimate of reliability in relation to "the single thing" (Cortina et al., 2020).

In the scenario of increased factorial complexity when multiple dimensions are present within a scale or a set of items, alpha is much more likely to be inflated due to higher interrelatedness among multidimensional items (which might not even measure the same thing) (McNeish, 2018). Meanwhile, there is a strong likelihood that the value of omega total is higher than that of hierarchical omega. We can see, from the above definitions of omega total and hierarchical omega, that hierarchical omega will never be greater than omega total and that they will only be equal when the scale is unidimensional. In other words, hierarchical omega will not be inflated by group factors or increased factorial complexity while omega total can be higher because both general factor and group factors are involved in the computation process.

Although we acknowledge the fact that researchers cannot always ensure the unidimensionality of their scales, we recommend that, out of methodological rigor, one should conduct factor analysis to resolve the issue of possible multidimensionality before rushing into examining reliability (Cortina et al., 2020; Kelley & Pornprasertmanit, 2016; McNeish, 2018). One way to remove multidimensionality is to separate sub-scales and report and calculate reliability estimates, respectively (McKay & Plonsky, 2021). A multidimensional scale easily leads to ambiguity about the focal

construct to be measured, hence weakening the clarity of study design. It can also possibly mislead researchers to make interpretations about a general factor when the results might be impacted by group factors, covering "the truth" in the population that is supposed to be found out.

In terms of interpreting omega coefficient estimates, therefore, we should proceed with great caution when we notice a combination of a much higher omega total estimate and a low hierarchical omega estimate. A low hierarchical omega value indicates small loadings on the general factor, and that the overarching construct is relatively weak and in fact cannot dominate all items. However, the contributions from sub-constructs are higher when the omega total value is high, which can interfere with results (Cortina et al., 2020). In this case, we should revisit all the items to ensure that the target construct is the only focus and that there is nothing else introducing noise into the data (i.e., unidimensionality). What we look for should be a high hierarchical omega estimate, implying that all items share a great deal in common, and similar omega total and hierarchical omega estimates, suggesting unidimensionality. Plonsky and Derrick (2016) reported a median instrument reliability estimate of .82 and provided a practical benchmark for applied linguists in their meta-analysis of reliability coefficients in L2 research. Meanwhile, they pointed out that the number of .82 should not be seen as a fixed threshold to decide whether a reliability coefficient is acceptable or not; instead, it is recommended that researchers interpret their own findings in comparison to this benchmark and validate the findings by considering the reliability estimates in similar contexts. We also hope to remind readers of the current over-prevalence of alpha estimates used in our field, which make up the most instrument reliability index in this meta-analysis. Therefore, caution should be exercised when omega coefficients are used and compared with the benchmark which may not apply to omega coefficients.

Another strategy for interpreting observed reliability estimates involves turning to the growing body of reliability generalization

meta-anayses (RGMs). Such studies, as the name implies, aggregate reliability estimates that pertain to a particular domain, variable, and/or scale. For example, in their meta-analysis of the relationship between L2 reading performance and working memory, In'nami et al. (2022) found the mean reliability estimates for working memory tasks with and without a processing task to be .81 and .60, respectively. Other recent examples of RGM in applied linguistics include Zhao & Aryadoust, in press, and Kostromitina & Plonsky, 2022).

Both omega total and hierarchical omega estimates can be calculated in the R software environment (R Core Team, 2023) with the psych package (Revelle, 2023), and omega total can be calculated using JASP. We recommend that interested readers learn more about the two coefficients from Cortina et al. (2020), Kelley and Pornprasertmanit (2016), and McNeish (2018) for both technical details and user-friendly explanations. (For an alternative perspective on the relative merits of these indices, see Raykov & Marcoulides, 2019).

4. Correcting for Attenuation

As mentioned in a previous section, low instrument reliability will necessarily reduce the magnitude of an observed correlation. The attenuation of observed effects is almost always present but is very rarely accounted for in applied linguistics research, which tends to mistakenly assume perfect reliability. If the relationship is underestimated, the risk of Type II error will also increase as real and statistically significant relationships are less likely to be found in the presence of measurement error. As the ultimate aim of research is to capture the "real" relationship in the population, and measurement error is inherently inevitable, researchers can and should consider correcting for low reliability and consequent attenuation (Osborne, 2003).

The method of correction for attenuation is cognitively clear and statistically simple. To start with, let us revisit the concept of reliability, which is the consistency that an instrument shows when measuring a given construct. In other words, reliability is conceptually similar to the correlation between the observed results from an instrument and the true results in an ideal scenario. If there is always some distance from the observed results and the true results (i.e., measurement error), the maximum possible correlation observed between two constructs that we want to examine is always limited by the reliability of instruments. The attenuated correlation is described in Equation 1, where r_{XY} is the attenuated correlation coefficient between variables X and Y, and R_{XY} is the true correlation coefficient. The degree to which the true correlation is attenuated is the geometric mean of reliabilities, $\sqrt{reliability_X reliability_Y}$.

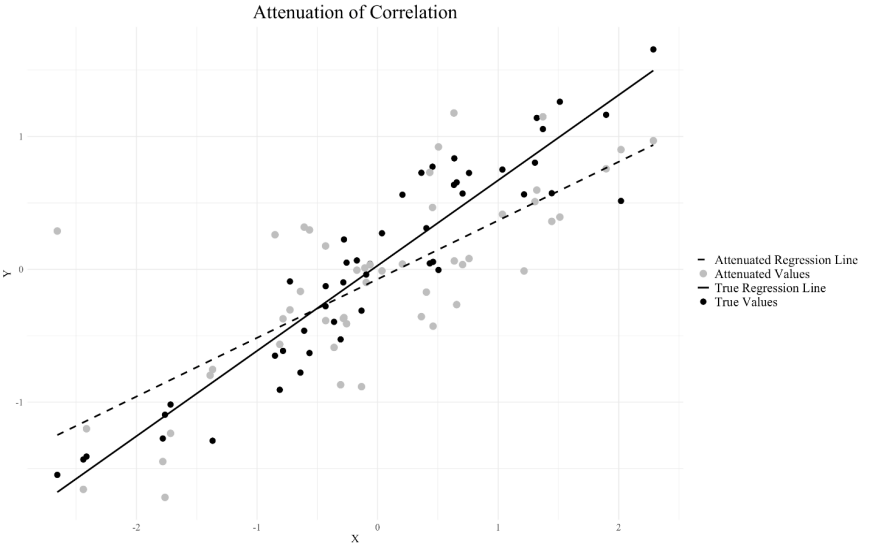
$$r_{XY} = R_{XY}\sqrt{reliability_X reliability_Y} \quad (1)$$

Figure 1 also demonstrates the attenuation effect. The black dots are true values, and the solid line is the true regression line, representing the actual (unattenuated) relationship between variables X and Y. In contrast, the gray dots represent attenuated values, which are the observed datapoints that have been affected by measurement error. The dashed line is the attenuated regression line and the best fit for attenuated values. The difference between the true and attenuated regression lines presents how imperfect reliability reduces the size of the observed correlation.

Fortunately, we can correct for this attenuation in observed effects. Specifically, we can calculate the corrected correlation coefficient by transforming Equation 1 to Equation 2.

$$R_{XY} = \frac{r_{XY}}{\sqrt{reliability_X reliability_Y}} \quad (2)$$

Figure 1. Attenuation of Correlation



Take the constructs of L2 motivation and L2 Willingness to Communicate (WTC) as an example. Suppose that the reliability estimates calculated for the two scales measuring L2 motivation and WTC are .85 and .82 respectively, and that the observed correlation coefficient, r_{XY}, is .55. With the help of Equation 2, we can correct for attenuation and compute the true correlation coefficient, R_{XY}, which is .66. To help readers better understand the impact of measurement error and ensure transparency, both attenuated and corrected correlation coefficients should be reported.

Since the goal of corrections for attenuation is to better capture a "true" relationship, we should also bear in mind that overcorrection can also happen and needs to be prevented (Osborne, 2003). Overcorrection is most likely to occur when reliability is underestimated, which, again, requires researchers to make informed decisions on choosing appropriate reliability estimates.

5. Conclusion

As we hope to have made clear in this article, an understanding of reliability is central to our ability to produce and interpret quantitative research in the field. However, our collective understanding of reliability is limited both at the conceptual and technical levels. Our goal in this article was, therefore, to provide an overview of some of the major issues at play as well as to highlight paths forward in the ways that we employ reliability estimates. We look forward to seeing the field's use of reliability improve and advance in ways that further our understanding of language learning, teaching, assessment, and usage.

References

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research*, *23*(6), 727–744. https://doi.org/10.1177/1362168818767191

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology*, *105*(12), 1351–1381. https://doi.org/10.1037/apl0000815

Gonulal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *ITL - International Journal of Applied Linguistics*, *168*(1), 4–32. https://doi.org/10.1075/itl.168.1.01gon

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods,

recommendations, and software for composite measures. *Psychological Methods*, *21*(1), 69–92. https://doi.org/10.1037/a0040086

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition, 44*, 886-911. https://doi.org/10.1017/S0272263121000395

Larson-Hall, J., & Plonsky, L. (2015). Reporting and Interpreting Quantitative Research Findings: What Gets Reported and Recommendations for the Field. *Language Learning*, *65*(S1), 127–159. https://doi.org/10.1111/lang.12115

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, *48*(2), 360–388. https://doi.org/10.1002/tesq.128

McKay, T., & Plonsky, L. (2021). Reliability analyses: Estimating error in L2 research. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 468–482). Routledge.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Osborne, J. W. (2002). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research, and Evaluation, 8*(1), 11. https://doi.org/10.7275/0k9h-tq64

Plonsky, L., & Derrick, D. J. (2016). A Meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*(2), 538–553. https://doi.org/10.1111/modl.12335

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*, 200-210. https://doi.org/10.1177/0013164417725127

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.

Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, *45*(5), 1427–1455. https://doi.org/10.1017/S0272263122000560

Zhang, P., & Han, C. (2024). Examining statistical literacy, attitudes toward statistics, and statistics self-efficacy among applied linguistics research students in China. *International Journal of Applied Linguistics*, *34*(2), 433–449. https://doi.org/10.1111/ijal.12500

Zhao, H., & Aryadoust, V. (in press). A meta-analysis of the reliability of second language reading comprehension assessment tools. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263124000627

Shujing Zhao is a Ph.D. student in Applied Linguistics at Northern Arizona University. She received her master's degree from the University of Pennsylvania in TESOL. Her research focuses on (instructed) second language acquisition, with an emphasis on individual differences and task-based language teaching, and research methods.

Luke Plonsky (PhD, Michigan State) is Professor of Applied Linguistics at Northern Arizona University and Honorary Professor at University of St. Andrews (Scotland). His work, focusing primarily on SLA and research methods, has appeared in over 100 articles, book chapters, and books. Luke is Editor of SSLA, Managing Editor of Foreign Language Annals, and Founding Editor of Applied Linguistics Press. He has held faculty appointments at Georgetown University and University College London, and was a Fulbright Scholar in Spain in 2021.

First version received: September, 2024
Final version accepted: October, 2024