Received: March 10, 2012 *Accepted: March 20, 2012*

Theoretical Misconceptions and Misuse of Statistics: A Critique of Khodadady and Hashemi (2011) and Some General Remarks on Cronbach's Alpha

Rüdiger Grotjahn[1]

Abstract

This article comments on theoretical misconceptions and misuses of statistics in Khodadady & Hashemi's (2011) paper "Validity and C-Tests: The Role of Text Authenticity". Firstly, it is pointed out that the pertinent C-Test literature is not adequately dealt with. Then, it is argued that the authors misconstrue the notion of the C-Test when they apply the term to a single (longer) C-Test text such as their AC-Test (Authentic C-Test). Subsequently, it is shown that there are serious flaws in the data analysis and interpretation. Here, the main focus is on local item dependence, which is not taken into account, and misconceptions with regard to Cronbach's Alpha, an issue of relevance to a wider audience.

***Keywords**: C-Test, Authenticity, Reliability, Cronbach's Alpha, Guttman's Lambda2, Local Stochastic Dependence, LID*

1. Introduction

Having been involved for almost thirty years in the study of C-Tests, and especially in construct validity research, I was delighted that the first issue of the new *Iranian Journal of Language Testing* contained an article devoted to the important question of the role of text authenticity in the validity of C-Tests. However, on reading the article authored by Khodadady and Hashemi (henceforth K&H), I realized that it was severely flawed. In the following, I will comment on these flaws. The main focus will be on the use and interpretation of Cronbach's alpha. Since alpha is the reliability coefficient most often applied in language testing, this part of my commentary will be of interest not only to C-Test researchers but also to a wider audience of language testers. I will start with K&H's review of the C-Test literature.

[1] *Seminar für Sprachlehrforschung (Department of Foreign Language Research), Ruhr-Universität Bochum, Bochum, Germany. Email: ruediger.grotjahn@ruhr-uni-bochum.de*

2. Major Flaws

2.1 Incomplete Coverage of the Pertinent Literature

K&H's literature review is incomplete. Contrary to Khodadady's (2007) claim that there is a "lack of research on C-Tests", C-Tests are among the best researched testing instruments (cf. Eckes & Grotjahn, 2006; Grotjahn, 2010, 2012). My updated C-Test bibliography, to be published next year, contains more than 300 items, including numerous publications in English. Furthermore, some of the pertinent publications not taken into account by K&H are easily available at no cost at www.c-test.de (by clicking "Originalia"). Moreover, several publications mentioned are not adequately dealt with by the authors. I will provide concrete examples later on.

2.2 Misconstrued Notion of "C-Tests"

In their review of the literature, K&H describe six advantages of C-Tests over cloze tests, quoting Klein-Braley (1997) who, together with Ulrich Raatz invented the C-Test principle. According to Klein-Braley (1997, p. 65) and numerous other C-Test researchers, one of the major advantages of C-Tests over cloze tests is:

> Because the C-Test consists of a number of different texts the sampling of content classes is better. Examinees who happen to have special knowledge in certain areas no longer have substantial advantages over other examinees.

This statement implies that C-Tests always consist of several texts and that therefore a single long C-Test text with 180 gaps such as K&H's authentic C-Test (AC-Test) is not a genuine C-Test. The statement also implies that K&H should not refer to Klein-Braley's individual C-Test texts as C-Tests (which is what they do).

Furthermore, when advocating their authentic C-Test, K&H do not mention that text authenticity has been addressed by various C-Test researchers, including Klein-Braley (1997, p. 64), who states in her description of the C-Test principle: "In short, carefully selected, *preferably authentic* texts we delete the second half of every second word beginning from word two in sentence two" (my emphasis). Although K&H on p. 31 explicitly refer to Klein-Braley (1997, p. 64), they only state "that between four to six carefully selected texts should be chosen", omitting the qualification "preferably authentic". In sum, K&H's use of an authentic text is not at all new and does not contribute to our knowledge about the validity of C-Tests. Most classic C-Tests consist of authentic texts (one exception being C-Tests for low-proficient learners). What is new, however, is K&H's use of *a single long* authentic text, calling this a C-Test. However, using only one text can lead both to content underrepresentation (e.g., with regard to lexis) and to (severe) bias and unfairness (cf. the quote from Klein-Braley), and, as a consequence, can jeopardize construct validity. In a standard C-Test, there is a chance that bias in one text may cancel out bias in another text. In addition, if we statistically check for differential item functioning (see Zumbo, 2007 for a brief overview of DIF) and detect that a text is severely biased with regard to gender for example, we must eliminate only the corresponding text and do not have to discard the whole measurement instrument as in the case of K&H's AC-Test. Other psychometric problems stemming from the use of only one text will be discussed below.

In addition to authenticity, K&H base their justification of the AC-Test on the following argument: They claim that Klein-Braley's (1997) suggestion that C-Test texts should be pre-piloted with native speakers or teachers of the corresponding language before administering them to the target population, "make the development of C-Tests very demanding if not too cumbersome, especially within a foreign language context where finding cooperative native speakers is too difficult, if not virtually impossible" (p. 31). It could be objected that in most contexts at least a few language teachers could be found who are willing to spare some of their time. Furthermore, as has been pointed out by several authors (again including Klein-Braley), even unpiloted C-Tests often prove to be psychometrically very robust when administered for the first time. A more substantial objection, however, is that Klein-Braley's suggestion that C-Tests be pre-piloted with native speakers or language teachers and also piloted with an adequate sample of language learners, applies equally to the AC-Test proposed by K&H, and, in my view, even more strongly.

2.3 Data Analysis and Interpretation

K&H start their data analysis by estimating what they call the "internal validity" of their tests (a standard C-Test taken from Klein-Braley, 1997, the AC-Test, and a retired version of the TOEFL).[2] In this context, they use the term internal validity as an antonym of external validity (as measured, for example, by the correlation between the C-Test and the TOEFL). This use of the term is rather misleading, since internal validity is a well-defined term in the literature on research methodology, designating the absence of confounding variables.

In order to evaluate the "internal validity" of the standard C-Test and their own AC-Test, K&H calculate difficulty values and discrimination indices (point biserial correlations) for the items of the tests, treating each gap as an item. Difficulty values falling within the range of .25 to .75 and discrimination indices of .25 or higher are considered acceptable by K&H. These analyses as well as their results are surprising in several respects. Firstly, in the standard C-Test, item difficulties range from .37 to .73 (Mean = .54, SD = .08), whereas in the AC-Test difficulties range from .16 to .91 (Mean = .54, SD = .17). The small range of item difficulties in the standard C-Test stands in sharp contrast to my own extensive data sets and also to the data reported by Klein-Braley (1996), Jafarpur (1999) or Babaii & Ansari (2001). For example, in Babaii & Ansari (2001, Appendix B), many items (particularly easy ones) lie outside K&H's difficulty range, although mean item difficulty was .55 and thus (almost) the same as in K&H's standard C-Test. This is all the more surprising as K&H's subjects were junior (n = 88) and senior (n = 47) undergraduate students majoring in English Language and Literature who should be relatively proficient in English. Since the words to be reconstructed include very simple (structure) words, one would expect some very high facility values (exceeding .73). Similarly, there should be at least some items with a difficulty index below .37. Incidentally, this expectation is met in the AC-Test for which K&H report five extremely difficult and 17 extremely easy items (my own recount yielded a total of 24 extremely easy items). In the light of this, I wonder whether the difficulty values in the standard C-Test are correctly calculated.

[2] According to Klein-Braley (1997, p. 66), her C-Test consists of four texts, each with 25 gaps. However, in the C-Test reproduced on pages 79f., text 2 has only 24 gaps. K&H's analyses are based on this C-Test containing 99 gaps.

Secondly, K&H report 46% well-functioning items for the standard C-Test and 54% for the AC-Test. They conclude that both the standard C-Test and the AC-Test enjoy acceptable internal validity, but that "the AC-Tests are superior to their standard counterparts in terms of their internal validity" (p. 36). This conclusion is flawed in several respects. (a) The authors should not generalize from the single AC-Test investigated to "AC-Tests and their counterparts." (b) The reported number of well-functioning items in the AC-Test is not correct (at least according to Table 4). It appears that K&H have taken into account only the values for the discrimination index. There are, however, an additional 12 items which are acceptable with regard to discrimination but not in terms of difficulty. If these items are also taken into account, the number of well-functioning items decreases to 47% and is thus (almost) the same as in the standard C-Test. (c) K&H provide no justification why internal validity is acceptable when around 50% of the difficulty and discrimination values fall outside their posited ranges. (d) K&H do not take into consideration that if the number of well-discriminating items (gaps) in an AC-Test is higher than in a standard C-Test, this may be the result of the greater amount of local item dependence (LID) within the AC-Test, compared to the standard C-Test consisting of several texts. Thus, if there is LID, high discrimination (calculated on the basis of individual gaps) does not necessarily mean better measurement. However, even if there is no LID, evaluating a measurement instrument primarily on the basis of item discrimination is not a good strategy since striving for high discrimination may result in construct underrepresentation and hence in lower validity. I shall return to the problem of LID when dealing with the issue of reliability estimation.

In order to evaluate the precision of measurement and reliability, K&H calculate standard deviations and Cronbach's alpha for each C-Test text, the C-Test total score, the AC-Test, each TOEFL subtest, and the TOEFL total score. As the AC-Test had the highest standard deviation followed by the TOEFL and then the standard C-Test, they conclude that "the AC-Test distinguishes among the test takers better than both the TOEFL and the standard C-Tests" (p. 34). In addition, since the reliability coefficient was higher for the AC-Test than for the standard C-Test (.92 vs .82), K&H concluded that "the AC-Tests are superior to their standard counterparts in terms of their reliability" (p. 34). They then make the following comment (the quote is verbatim):

> While it is argued that higher standard deviation (SD) and reliability coefficient of the AC-Test is due to its length, it does not necessitate standardizing the SDs as suggested to compare them with each other for two reasons. First, SDs are standardized by their very nature and secondly there is no theoretically sound basis to establish a cut off number for the items comprising the C-Tests, i.e., 100, as Klein-Braley (1997) did. It is, in fact, argued in this paper that the inclusion of more items in the AC-Tests provides a more reliable and valid measure of test takers' ability… (p. 35)

These conclusions are flawed for several reasons. Firstly, standard deviations are not "standardized by their very nature". They clearly depend on the range of the scale and in comparing standard deviations, one has to take this fact into account. Therefore the conclusion that the AC-Test distinguishes best among the test takers, because it has the highest standard deviation, is not sufficiently substantiated and, as will be shown, incorrect.

Secondly, with regard to the reliability of C-Tests, Klein-Braley (1997, p. 63) proposes that to be sufficiently reliable C-Tests should comprise *at least* 100 gaps. (A practical advantage of

having exactly 100 gaps is that the raw scores do not have to be converted into percentages.) Thus, 100 gaps is not a fixed cut-off, as erroneously argued by K&H, and there are C-Tests with less and also with more than one hundred gaps, depending on the desired level of reliability and the stakes involved in the examination (for a medium-stakes online German C-Test with 8 texts, each containing 20 gaps, see www.ondaf.de). If K&H's standard C-Test had the same length as their AC-Test, its reliability would probably be very similar to that of the AC-Test.[3]

Thirdly, K&H's reliability estimation for the C-Test and the AC-Test is flawed since the authors calculate Cronbach's alpha on the basis of the individual gaps. The same holds true for earlier publications by Khodadady (cf., e.g., Khodadady, 2007). As Cronbach's alpha is widely used in language testing, and as K&H's paper is a case in point for the "misunderstanding and confusion" with regard to alpha discussed in a recent article by the psychometrician Klaas Sijtsma (cf. Sijtsma, 2009a, p. 107), I will deal with this issue in some detail.

Cronbach's alpha as well as other reliability coefficients based on a single test administration assumes that the items are locally stochastically independent. If this assumption is not met, that is if there is local item dependence (LID) and correlated errors, reliability[4] tends to be overestimated (i.e, alpha is spuriously inflated) and estimates of item difficulty and person ability as well as decisions based on C-Test scores might be severely biased (cf. Sijtsma, 2009a, 2009b; Zhang, 2010). It has been pointed out repeatedly in the C-Test literature (including no-fee open access internet publications) that there is LID in C-Tests, stemming from at least two sources, namely passage dependency and item chaining, the latter meaning that reconstructing a mutilated word correctly or incorrectly may affect the reconstruction of other words in the same text (see, e.g., Grotjahn, 1987).

Several approaches have been used to tackle the problem of LID in C-Tests. The standard approach consists of collapsing the items in each C-Test text into one polytomous super-item or testlet by summing up the scores. These super-item scores (testlet scores) are then used for estimating the reliability by means of Cronbach's alpha or another reliability coefficient based on classical test theory such as Guttman's lambda2. Alternatively, the polytomous scores are used for further analysis based on item response theory (IRT) (for IRT approaches in C-Test research cf., e.g., Baghaei, 2010; Eckes, 2010, 2011; Eckes & Grotjahn, 2006). Note that if the C-Test texts differ considerably with regard to the number of deletions, Cronbach's alpha will severely underestimate reliability. In this case, a generalization of alpha can be used, namely Raju's beta coefficient (cf. Grotjahn, 1987, p. 229). However, since IRT approaches also become more complicated when the number of gaps is not the same for each text (cf. Baghaei, 2010; Eckes, 2010, 2011), it is advisable to have the same number of gaps in each text of a C-Test.

However, when constructing polytomous super-items from the individual item scores, part of the information carried by the individual item scores is lost. To avoid this loss of information, it

[3] If we use the well-known Spearman-Brown prophecy formula for calculating the effect of extending the standard C-Test from 99 to 180 items, reliability raises from .821 to .893, the latter value being very close to K&H's .921 for the AC-Test.

[4] There are various conceptions of reliability. Sijtsma (2009a, 2009b, 2012) discusses reliability in terms of repeatability of individual test performance and argues that "reliability estimates based on a single test administration, like alpha, may not convey much information about the accuracy of individual test performance" (Sijtsma, 2009a, p. 108). Sijtsma (2009a, p. 108) further argues that alpha is "misunderstood as a measure of internal consistency" and convincingly demonstrates that "alpha does not convey information about the internal structure of the test" (i.e. with regard to dimensionality).

has also been proposed to model the dependencies within a C-Test using IRT. There are several IRT models which can take item dependencies into account (cf. Eckes, 2011; Harsch & Hartig, 2010; Yen & Fitzpatrick, 2006). However, these complex IRT models require rather large sample sizes and are therefore not applicable to K&H's data.

Since K&H calculate alpha on the basis of the individual gaps and a single test administration, reliability is probably (severely) overestimated. This holds true especially with regard to the AC-Test and the individual C-Test texts, but less so with regard to the C-Test total score, since there is no item dependence across texts. As a further consequence, the reliability estimates for the various tests are only in part comparable. A better approach, at least with regard to the AC-Test and the individual C-Test texts, would have been to estimate reliability not on the basis of a single test administration, but by means of retesting (calculating the Pearson product-moment correlation between the scores from both administrations). It is somewhat surprising that K&H do not mention these problems at all, although the authors draw on several articles where local dependence and reliability estimation in C-Tests have been explicitly dealt with. A case in point is Babaii & Ansari (2001, p. 214), who report reliability estimates of .98 on the basis of individual gaps and .93 on the basis of super-items for a C-Test consisting of 8 texts with 20 gaps each.

In addition, when interpreting alpha, one must also take into account that alpha is a function of both the average inter-correlation among items and the number of items. Therefore, alpha tends to increase when the number of items increases, provided that the additional items correlate positively with the items already in the test. This is even true if the additional items have rather low discrimination values (e.g., because they partly measure something else). This is another reason why alpha values for tests with different lengths cannot directly be compared.

Furthermore, alpha is only then a point estimator of reliability if the items are essentially τ-equivalent, that is if the covariance is the same for each item pair as well as for all items and any independently measured variable (cf. Sijtsma, 2009a, p. 111). This also implies unidimensionality of the item set. However, if the items are only congeneric, i.e. one-dimensional,[5] rather than essentially τ-equivalent, which is usually the case with real-world test data, then alpha is only an estimator of a lower bound to the true reliability, and not even the best one. A better estimator is, for example, Guttman's lambda2, which is also available in SPSS (for lambda2 cf. Sijtsma 2009a and already Grotjahn, 1987). In addition, since alpha is also liable to sampling error, the lower bound may be both overestimated and underestimated.[6] Sijtsma (2012, p. 7) refers to the fact that values of alpha reflect both random measurement error and sampling error as "double stochasticity" and admits that "it is a difficult phenomenon to understand." Sampling error could be taken into account by constructing confidence interval around alpha or testing alpha for significance, which is however seldom done. In any case, small differences between alpha values should be interpreted more cautiously than K&H do, especially if sample sizes are small.

[5] These assumptions could be tested with the help of structural equation modeling (cf. Sijtsma, 2009a, 2009b).

[6] Underestimation of reliability can lead to an overestimation of correlations that are corrected for attenuation (i.e. for measurement error) and hence to wrong theory building.

3. Conclusion

Attempting to validate a measurement instrument, as the title of K&H's paper implies, means building a validity argument which consists of various pieces of evidence and counter-evidence. K&H have only sought to present evidence, while neglecting easily available counter-evidence. Moreover, the evidence presented is severely flawed. I can therefore only conclude that K&H's paper is definitely not an instance of good practice in language testing.

References

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System, 29*(2), 209-219.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling, 52*(3), 313-322 [http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2010_20100928/06_Baghaei.pdf].

Eckes, T. (2010). Rasch models for C-Tests: Closing the gap on modern psychometric theory. In A. Berndt & K. Kleppin (Eds.), *Sprachlehrforschung: Theorie und Empirie. Festschrift für Rüdiger Grotjahn* (pp. 39–49). Frankfurt am Main: Lang.

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling, 53*(4), 414-439 [http://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2011_20111217/02_eckes.pdf].

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290-325.

Grotjahn, R. (1987). How to construct and evaluate a C-Test: A discussion of some problems and some statistical analyses. In R. Grotjahn, C. Klein-Braley & D. K. Stevenson (Eds.), *Taking their measure: The validity and validation of language tests* (pp. 219-253). Bochum: Brockmeyer. Available at: [http://www.c-test.de/deutsch/index.php?lang=de&section=originalia]

Grotjahn, R. (Ed.). (2010). *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research*. Frankfurt am Main: Lang.

Grotjahn, R. (Ed.). (2012). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends*. Frankfurt am Main: Lang (to appear).

Harsch, C., & Hartig, J. (2010). Empirische und inhaltliche Analyse lokaler Abhängigkeiten im C-Test. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 193-204). Frankfurt am Main: Lang.

Jafarpur, A. (1999). Can the C-test be improved with classical item analysis? *System, 27*(1), 79-89.

Khodadady, E. (2007). C-Tests: Method specific measures of language proficiency. *Iranian Journal of Applied Linguistics, 10*(2), 1-26.

Khodadady, E., & Hashemi, M. (2011). Validity and C-Tests: The role of text authenticity. *Iranian Journal of Language Testing, 1*(1), 30-41.

Klein-Braley, C. (1996). Towards a theory of C-Test processing. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 3, pp. 23-94). Bochum: Brockmeyer. Available at: [http://www.c-test.de/deutsch/index.php?lang=de&section=originalia]

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*(1), 47-84.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika, 74*(1), 169-173.

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika, 77*(1), 4-20.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: American Council on Education/Praeger.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*(1), 119-140.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.