# *The International Journal of Language Learning and Applied Linguistics World*

(IJLLALW)

## (IJLLALW)

## Volume 2 (1), January 2013

**ISSN: 2289-2737**

IJLLALW

# TABLE OF CONTENTS

*The International Journal of Language Learning and Applied Linguistics World (IJLLALW)*
**Volume 2 (1), January 2013** **ISSN: 2289-2737**
**www.languagelearningworld.org**

# Authenticity and Sampling in C-Tests: A schema-Based and Statistical Response to Grotjahn's Critique

**Ebrahim Khodadady**
*Ferdowsi University of Mashhad*

## ABSTRACT

The paper aims to respond to Grotjahn's (2012a) critique of Khodadady and Hashemi's (2011) paper "Validity and C-Tests: The Role of Text Authenticity" by employing reduced redundancy (RR) and schema theories. It counter argues that developing conventional C-Tests on several short texts and modifying their contents do not render them "genuine" for RR has nothing to do either with the number and length of texts to be chosen or with mutilating a set number of words constituting those texts. Without acknowledging, however, the conventional C-Test designers resort to the macrostructural view of schema theory to justify measuring "special knowledges" assumed to be conveyed in several texts. They do, nonetheless, utilize its micro structures, i.e., their constituting words, when they mutilate every second word from the second sentence and onwards. Based on the RR and schema theory as well as the texts selected and the data presented by several authors, the points raised by Grotjahn are discussed and suggestions are made for future research.

**Keywords**: Schema theory, reduced redundancy, sampling, authenticity

## INTRODUCTION

Critiquing my joint paper "Validity and C-Tests: The Role of Text Authenticity" (Khodadady & Hashemi, 2011) [henceforth K&H] Grotjahn (2012a) announced that "it was severely flawed" (p. 12). Although the very time Grotjahn had spent on the critique necessitated appreciation, I found it rather surprising that he had not mentioned even a *single* strongpoint in the entire paper. As a language educator who has offered various courses in applied linguistics at both undergraduate and graduate levels in general and language testing in particular for over two decades to both native and non-native speakers of English, I have always done my best to help both language learners and teachers see the advantages of a research project, however flawed it might sound to them, before they focus on its possible disadvantages. Therefore, I appreciate Grotjahn's (2012) critical reading of my earlier paper as well, i.e., C-Tests: Method Specific Measures of Language Proficiency (Khodadady, 2007), which provides the necessary background for the points I will raise to address the *assumed* flaws. In contrast to Grotjahn, I will choose more *positive* headings in order not to activate unfavorable schemata in my reader's minds before they formulate their own on the basis of what they read. I will first focus on review of literature and then address the theoretical foundation of C-Tests and the authenticity of texts upon which they must be developed if they are accepted and employed as measures of language proficiency.

## REVIEW OF LITERATURE

Grotjahn (2012a) claimed that "K & H's literature review is incomplete" (p. 12). However, instead of referring to K&H to support his claim, he refers to Khodadady (2007). In addition to the fallacy of criticizing the content of one study on the basis of the content of another, what Grotjahn quotes from Khodadady is, unfortunately, removed from its context. Based on the decontextualised phrase "lack of research on C-Tests" quoted from Khodadady, he announces that "C-Tests are among the best researched testing instruments" (p. 21) and provides Eckes and Grotjahn (2006) and Grotjahn (2010, 2012b) as his chief references.

Since in his critique Grotjahn (2012a) might have unintentionally left out the paragraph preceding the quoted phrase, it is quoted below to provide the necessary context.

Although C-Tests were invented in 1981, they have received little attention in *English* language testing literature. For example, in his fairly comprehensive review of correlational studies conducted on C-Tests so far, Sigott (2004, pp. 61-65) could tabulate 28 among which only 11, i.e., 39%, have been in English. This is reflected in textbooks written for teacher training programs, e.g., Madsen (1983), Heaton (1988), Baker (1989), where C-Tests are not even mentioned. Similarly, there is no entry for C-Tests in the *Dictionary of Language, Teaching and Applied Linguistics* (Richards, Platt, & Platt, 1992). Bachman (1990), however, referred to C-Tests as variants of cloze tests in passing (Khodadady, 2007, p. 21).

Based on the reasons outlined above for the lack of long due original research papers written in *English*, which was originally italicized in the quoted paragraph to emphasize the variable of language in which the research projects need to be done *as well as* reported, Khodadady (2007) argued that

The overall public and expert inattention might be attributed partly to a lack of research on C-Tests and partly to their nature. Davies (1990) dubbed them 'a particular and rather recondite use of the cloze test' (p. 94) and thus obliged researches like the present one to contribute to those studies which have already shed some light on their internal, empirical and factorial validity (p. 21).

The first directly quoted paragraph preceding the phrase quoted in the paragraph above, i.e., a lack of research on C-Tests, is composed to draw the attention of testing experts and students alike to the very fact that there is still an urgent need to design and conduct research projects on C-Tests and report them in *English* so that scholars who write textbooks in English can include C-Tests as important, though method specific (Khodadady, 2007), measures of language testing. While I appreciate the attempt of designers of the site given by Grotjahn, i.e., www.c-test.de, to provide interested readers with open-access links on C-Tests, *even* that site falls short of providing diverse enough studies written in English to support his decontextualized argument. Out of 43 links, only 18 (41.9%) are in *English*, showing statistically that 58.1% of links are all in *German* not in *English*. Furthermore, C-Tests need to be developed in Arabic, Persian, and Turkish, to name a few. [This suggestion does not mean that I am unaware of the C-Tests designed in languages such as French and Japanese!]

***Prescriptive Vs. Descriptive Approach in Testing***

Grotjahn (2012) claimed that the K & H's C-Test developed on a single authentic text is not *genuine*. The claim is based on Klein-Braley's (1997, p. 65) belief that "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" (p. 65). According to Grotjahn (2012a) the belief "implies that C-Tests *always* [Italics added] 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" (p. 21). With due respect, I disagree with Klain-Braley's belief and what it implies to Grotjahn. What she says is not an *absolute* rule to be followed by test designers. I strongly believe that the time for prescribing certain rules to be followed by everyone is long over. As a matter of fact, science established itself as an indispensible tool to understand the functioning of various variables when it questioned the validity of some beliefs held by authorities.

It is, for example, argued in this paper that Klain-Braley (1997) contradicted herself unconsciously when she employed reduced redundancy (RR) as a theoretical rationale for C-Tests but followed schema theory in practice. While the former has nothing to offer as far as the number of texts employed in C-Tests is concerned, the latter does address and necessitate the selection and inclusion of various texts by resorting to its macro structural approach. To support the argument, schema theory will be described, albeit briefly, and then C-Tests will be analyzed within an RR perspective.

***Macro and Micro Structural Views of Schema Theory***

According to schema theory, reading comprehension ability can be measured either macro structurally or micro structurally (Khodadady, 1997, 1999). A given text is viewed as a single schema or macro structure which requires "special knowledge" in Klein-Braley's (1997, p. 65) words or "a conventional knowledge structure that exists in memory" in Yule's (2006 p. 132) perspective. However, no one still knows, as Grabe (2002) put it, "how it would work for reading comprehension" (p. 282). Although cloze tests were developed originally as measures of readability, they were later employed in language testing as integrative measures of language proficiency assuming that test takers' would employ their language proficiency, not their "special knowledge" of *a* given text, to restore its deleted words on the basis of hypotheses they formulate as they read the text (Khodadady & Herriman, 2000). Klein-Braley, however, interpreted language proficiency as "special knowledge" of *a* given text and extended it to "*special knowledge**s***" of several texts. (Notice that the plural morph –**s** is deliberately added to knowledge by the present author.)

Although Klein-Braley (1997) subscribed to the macro structural view of schema theory when she emphasized "*special knowledge**s***" measured by several short texts, she violated one of its main principles when she developed her C-Tests on *general topics* assuming that these general topics are synonymous with *allegedly different* areas of human knowledge. In order to help the test takers activate their "special knowledge" of a certain text or schema test designers conventionally provide its title and leave its first sentence intact. However, Klein-Braley

employed Text 1, 2, 3 and 4 as the titles of the four piloted texts on which she developed her final C-Test. She did, nonetheless, leave the first sentences of these texts intact as the second principle followed by followers of macro structural/top-down approach of schema theory stipulates. (It must be emphasized that reduced redundancy cannot, by its very nature, say anything about providing the title and first sentences because its occurrence is based on randomness or probability. In natural settings any part of any text can go missing because of variables such as noise and distraction.)

Since there is no title for the texts employed by Klein-Braley (1997) to address the nature of "special knowledge," the titles of the texts selected by Babaii and Ansary (2001) are given here. These conventional C-Test designers have been chosen because Grotjahn (2012a) seemed to have agreed with their selection of texts and reliability analyses. The titles are "A slip of the Tongue", "the End of the World", "A 50 Percent Thief", and "Keep the Torch Burning". These topics deal, according to the present author, with the usual issues encountered by almost all proficient English readers. It remains to be found out how Klein-Braley and Grotjahn would justify their claim that the "special knowledge" of a given test taker of C-Test developed on "A slip of the Tongue" will, for example, be different from, say, that of the second or third test taker?

Similarly, it remains to be explained by Grotjahn how, say, the first test taker's "special knowledge" of "A slip of the Tongue" will be different from his knowledge of "The End of the World" or "A 50 Percent Thief". This assumption violates Spolsky's (1973) insistence on a *single* "knowledge of the language" (p. 7) based upon which a test taker can restore the missing parts of a message. In other words, *a single knowledge of the language* will be enough to restore the missing parts of *any* given text and thus assuming the existence of "*special knowledges*" on the part of test takers violates the principle upon which C-Tests are designed.

Interestingly enough Grotjahn (2012a), however, objected to K&H's use of using ***C-Tests*** in plural and emphasized that Klein-Braley used its singular form, i.e., ***C-Test***. As discussed previously, the use of ***C-Test*** is not only misleading but also flies against the results reported by Klein-Braley (1997) herself as shown in Table 1. In describing the table she wrote, "the reliability coefficients have been calculated for the *individual tests* [italics added] using KR-21" (p. 67). As can be seen, the descriptive statistics of C1, C2, C3 and C4 are reported as *individual tests*. Since Klein-Braley could not solve the self-created problem of using the singular C-Test for each of its four constituting tests, she employed C1, C2, C3 and C4. Khodadady (2007) simply referred to them as C-Test 1, C-Test 2, C-Test 3 and C-Test 4, respectively, and treated them as *individual tests* as Klein-Braley (1997, p. 67) herself did.

***Table 1: Basic test statistics for the tests of reduced redundancy (Klein-Braley, 1997, p. 67)***

| Test | Mean | SD | *P* | $rtt_{KR-21)}$ | $rtt_{(DELTA)}$ |
|---|---|---|---|---|---|
| CLOZE1 | 6.38 | 2.91 | .32 | .51 | .71 |
| CLOZE2 | 4.34 | 2.58 | .22 | .52 | .58 |
| C1 | 15.15 | 4.28 | .61 | .70 | .68 |
| C2 | 10.69 | 5.85 | .43 | .86 | .66 |
| C3 | 14.86 | 4.98 | .60 | .79 | .53 |
| C4 | 10.42 | 4.71 | .42 | .76 | .45 |
| MC1 | 16.32 | 3.31 | .63 | .46 | .55 |
| MC2 | 13.47 | 3.35 | .56 | .49 | .61 |
| CE1 | 14.10 | 4.89 | .56 | .72 | .56 |
| CE2 | 8.64 | 5.45 | .35 | .84 | .45 |
| DICT | 31.36 | 12.54 | .62 | .94 | .71 |
| DELTA | 83.40 | 22.07 | .56 | .93 | 1.00 |

It is further argued in this paper that the assumed "special knowledge" of a given text is based on macro structural school of schema theory adopted and operationalized by the designers of earlier IELTS modules (e.g., Clapham, 1996, Kelly, 1978). The assumption was that since the background knowledge required for understanding a given field such as humanities is different from another field such as engineering, the test takers wishing to continue their academic studies in engineering in English would be at a loss if their reading comprehension texts were chosen from humanities. What Klein-Braley (1997) and other designers of conventional C-Tests such as Babaii and Ansary (2001) did not notice was that neither titleless and numbered texts such as 1, 2, 3 and 4 nor texts dealing with general topics such as "a slip of the Tongue" and "the End of the World" are field-dependent and thus could not represent various types of "special knowledge". Even if they did, studies after studies showed that proficient test takers having "special knowledge" of a given field did not necessarily perform significantly differently on the tests developed in their field than those possessing "special knowledge" in a different field.

***C-Tests and Reduced Redundancy***

Since providing an operationalized definition of "special knowledge" in terms of schema theory to establish C-Tests as macro structural measures of language proficiency is too difficult, if not impossible, Klein-Braley (1997) employed RR as a viable rationale to provide C-Tests with a theoretical foundation. Similar to her, Spolsky (1973) is quoted below in order to find out whether it can be applied to conventional C-Tests.

The non-native's inability to function with reduced redundancy, evidence that he cannot supply from his knowledge of the language the experience on which to base his guesses as to what is missing. In other words, the key thing missing is the richness of knowledge of probabilities - on all levels, phonological, grammatical, lexical, and semantic - in the language (p. 17)

According to Klein-Braley (1997), Spolsky assumed that "knowing a language certainly involves the ability to understand a distorted message, to make valid guesses about a certain percentage of omitted elements" (p. 47). As can be seen, there is no indication of "special knowledge" of the distorted message neither in Spolsky's nor in Klein-Braley's quoted claims. Klein-Braley does, however, provide a lengthy review of literature dealing with cloze test as measures of RR without any indication of how "special knowledge" relates to RR and whether she has borrowed the concept from top-down models of reading or schema theory.

As discussed before, choosing a number of texts for the development of C-Tests is based on macro structural view of schema theory. If test takers coming from diverse fields such as humanities and engineering are going to take them as valid measures of language proficiency then choosing representative texts from their respective fields would be justified. However, it does not apply to the tests developed by Klein-Braley (1997) because they are not developed on any "special knowledge" which might be known to certain test takers, say those of humanities, and stay unknown to others, say engineering. It is, therefore, suggested that instead of choosing a number of short texts dealing with general topics, more C-Tests be developed on single authentic texts written for the literate English speaking public as K&H did. It is also suggested that C-Test

items be developed on first sentences to find out whether their being kept intact has any significant effect on test results.

***Randomness and Reduced Redundancy***

This paper attempts to show that RR has nothing to do with "special knowledge" upon which Grotjahn (2012a) questioned the development of an authentic (A)C-Test on a single text. It also tries to show that neither cloze tests nor C-Tests comply *fully* with the stated principle of RR as conceived by Spolsky. The concept of noise as the main cause of distortion in native speakers' reception of messages in real life is based on *probability*, i.e., it occurs randomly and may affect any parts of an authentic text. Although Spolsky (1973) had observed the occurrence of RR, he remained quite vague in its description as far as the present author's knowledge and experiences allow. Spolsky believed, for example, RR occurred "on all levels, phonological, grammatical, lexical, and semantic - in the language" (p. 17). It remains to be researched, for example, whether and how *semantic* is affected by RR. To begin with semantic is an abstract concept as is phonology. How would Spolsky himself and other believers in RR translate these ***abstract concepts*** into describing ***concrete missing parts*** of a distorted text?

Klein-Braley (1997) seems to be the first who translated Spolsky's belief into practice *quantitively* by announcing that "knowing a language certainly involves the ability to understand a distorted message, to make valid guesses about *a certain percentage* [italics added] of omitted elements" (p. 47). Along with her colleagues she developed C-Tests by mutilating *at least* 100 words constituting the second and subsequent sentences of texts. Grotjahn (2012a) offered "a practical advantage" as a rationale saying that "the raw scores do not have to be converted into percentages" (p.24). The present author could not, however, make out how "a practical advantage" can be used as ***a theoretical basis*** as K&H wrote, "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" (p. 35).

The concept of randomness in RR is emphasized in this paper because it embodies several cardinal variables playing significant roles in language testing. Surprisingly however, Spolsky (2001) revealed his unfamiliarity with these variables when he preferred C-Tests over cloze tests simply because the results obtained on the former do not support his *vague* theory of RR. According to him,

By omitting words, which are *linguistic* [italics added] elements with certain properties a cloze test was biasing itself to testing certain areas of language … the technique she [Klein-Braley] proposed as an alternative, the C-Test, used half words. A half word is much less linguistic - not a discrete item - and so much more information theory-oriented and integrative. Essentially, a C-Test was much closer to a noise test in the *randomness* [italics added] of the reduction of redundancy and so a purer example of an integrative rather than a discrete item test (p. 7).

What Spolsky (2001) stated in the quoted block above is self-condemning for several reasons/variables. *First*, how can deleting complete word be *linguistic* and *non-random* but their mutilation be *integrative* and *random*? If we consider integrativeness as an indispensible part of context, then the opposite will hold true. While it is impossible to restore any omitted word without having access to its context, Khodadady (2007) showed that out of 99 mutilated words comprising Klein-Braley's (1997) C-Tests, 11 functioned quite well when they were removed from their textual context and presented as single mutilated words to be restored on the basis of

directions given in C-Tests. In other words, test takers could restore eleven percent of items on the C-Test without having any context!

Secondly, RR cannot specify which words/items, i.e., every second word, must be mutilated if they are chosen *randomly*, hence Spolsky (2001) contradicted himself by claiming that C-Tests are "much closer to a noise test in the *randomness*" (p. 7). Thirdly, noise can affect all textual units, e.g., words, phrases and clauses, and the advocates of RR need, therefore, to justify the nature of the missing items in terms of the "phonological, grammatical, lexical, and semantic" levels specified by Spolsky (1973). Fourthly, RR cannot endorse leaving the title and first sentence of a given text intact because these two textual units might also go missing when it takes place in reality. And finally noise distorts *authentic* texts which are produced and processed for *real* purposes in *real* places at *real* times and for *real* purposes.

***Text Authenticity and C-Tests***

This paper supports Khodadady (2007) and K&H' view that authentic texts have ***not*** been used in the development of conventional C-Tests and questions Grotjahn's (2012a) decontextualized quotations of K & H"s sentences in order to prove that the development of C-Tests on authentic texts is endorsed and brought up by Klein-Braley (1997) *first*. He writes

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". (p. 21)

Grotjahn (2012a) quotes both K & H and Klein-Braley in an ambiguous way. His manner of quotation leads readers to the conclusion that K & H deliberately omitted the phrase "preferably authentic" from the sentence he quotes from K & H. In other words, the quoted *that clause* belongs to K & H, i.e., "that between four to six carefully selected texts should be chosen" (p. 31) whereas the phrase "preferably authentic" belongs to Klein-Braley (1997, p. 64). The following quotation provides the original context to which K & H referred to *indirectly*.

A number of texts, usually between four and six, are put together to make a C-Test. Because of problems with text difficulty, usually overestimated by the test constructor (cf. Klein-Braley, 1985b; 1994), one should begin with more texts than will be finally needed. The texts are ordered intuitively according to difficulty (Klein-Braley, 1997, p. 65).

Klein-Braley (1997) did suggest the selection of "preferably authentic" (p. 64) texts. However, she did not provide her readers with any specific definition of authenticity. Neither did she supply them with any references to verify her suggestion. For example, nobody knows what sources she used to select Texts 1, 2, 3 and 4 to develop her C-Tests from. In contrast, K&H employed "why don't we just kiss and make up" (Dugatkin, 2005) published in *NewScientist* magazine whose authenticity can be verified by all interested readers. They chose this magazine because its articles are "more academic than … articles in quality newspapers" (Clapham, 1996, p. 145) and they provide standard scientific texts for public readership.

Not only did Klein-Braley (1997) provide no references to find out what sample authentic texts she had preferred but also she believed that "no language test is *authentic*" (p. 48). She argued that "normal language is not produced in order to be assessed." The present author, however, argues that English writers produce normal/authentic texts to be read. *Whatever texts which are produced to be read by the literate public are authentic*. This argument was employed to question the construct validity of the TOEFL by highlighting the fact that its reading comprehension texts are written by language testing experts in order to *test* reading comprehension ability (see Khodadady, 1997). Since the texts upon which the TOEFL is designed are not *authentic*, i.e., *they are not written to be read*, it lacks construct validity. The same argument was employed by K&H to show that the four texts employed by Klein-Braley (1997) were not *authentic* because they were not written to be ***read***. If they were, she would have provided their sources or references.

***Sampling Authentic Texts for C-Tests***

According to Klein-Braley (1997), a number of texts need to be selected to develop C-Tests because their writers may face "problems with text difficulty" (p. 65) and thus may have to do away with some. Finding texts with appropriate difficulty is not only problematic for designing language proficiency tests such as C-Tests, it violates the principle of authenticity in that the authentic texts written to be read by literate public may have all levels of assumed difficulty. Furthermore, it poses a real problem which becomes more complicated and time consuming when the test designers realize that some of their chosen texts have functioned poorly in the pilot phase and they must, therefore, look for suitable substitutes. In Baghaei's (2008) words,

For developing a C-Test battery the number of the texts used should be more than the number required since even native speakers cannot obtain perfect scores (95%) on some texts. They believe that native speakers should perform perfectly on language tests. To what extent this view is credible is another issue (p. 33).

The selection of texts written to be *read* by the literate public not only ensures authenticity and dispenses with the necessity of administering C-Tests to *native speakers* but also relieves C-Test designers from looking for a number of texts with appropriate difficulty levels. K&H, therefore, chose a single text whose C-Test items functioned as well as conventional C-Tests developed on four texts as will be discussed shortly. The only reason Klein-Braley (1997) provided for the cumbersome and theoretically questionable process of choosing a number of texts instead of a single authentic text is her adamant attempt to sample texts addressing "*special knowledge**s***". This attempt is, nonetheless, misplaced because RR has little, if any, to do with text selection.

In addition to the fallacy of employing "special knowledge" as a synonym for a given area of knowledge such as humanities and engineering, the very necessity of choosing short texts in order to account for "*special knowledge**s***" fails to represent the types of texts proficient test takers are going to read when they enter higher education centers. If we take academic textbooks and articles as the most normal types of texts read by college and university students, none of these texts consist of a single paragraph! In other words, the C-Tests developed by Klein-Braley (1997) and her followers not only fail to accommodate authenticity in content and audience but also misrepresent academic texts as single short paragraphs!

Since Klain-Braely (1997) ***did not*** provide the sources of her texts so that their intended audiences could be objectively explored, those of Babaii and Ansari (2001) will be addressed as representative samples of conventional C-Tests. Their "eight excepts were taken from two ELT textbooks, viz *Practice and Progress* (Alexander, 1968) and *To start you practicing* (de Freitas,

1974)" (p. 217). Out of eight conventional C-Tests developed on eight texts ***written for teaching English***, only five were kept for final administration because three of them did not reveal Babaii and Ansari's expected item characteristic indices providing further evidence to support the earlier argument made in this paper, i.e., developing conventional C-Tests entails the cumbersome process of choosing more texts and trying them out in a pilot phase.

Babaii and Ansari (2001) distorted not only the texts of their C-Tests but also the title of the first book they gave as one of the two sources of their selected texts. The full title is not *Practice and Progress* but *Practice and Progress: An Integrated Course for Pre-Intermediate Students* which was published in 1967, not in 1968. As the original title implies, the content of the textbook was written for *teaching* English; therefore, the passages developed in the textbook were not *authentic* in that no proficient English user was supposed to *read* them for the sake of comprehension. Furthermore, the stated level of its would-be users is *Pre-Intermediate* which renders the C-Tests developed on the texts of this textbook questionable if not invalid in terms of their construct validity.

In addition to choosing the texts of an inappropriate proficiency level, Babaii and Ansari (2001) simplified the texts Alexander (1967) had already modified the texts to teach English to *Pre-Intermediate Students*. Alexander himself did not state where he got the passages from. However, he did declare that "each passage contains examples of the language patterns the student is expected to master" (xv) implying that they were particularly *written* for teaching purposes because their constituting number of words also varied according to the students' level of achievement. While the passages employed for Unit 1, for example, consisted of just one hundred words, they increased to 180 in Unit 4. The educational purposes of *artificially written* or *modified* passages is further emphasized in a section called "For Whom the Course is Intended", Alexander identifies four types of students among whom are "students in need of remedial work: e.g., … students who have begun English several times and never got beyond the point of no return" (p. xii)

From among the 26 passages presented in Alexander's (1967) Unit 4, Babaii and Ansari (2001) chose "A Slip of the Tongue" for inclusion in their C-Test. They refer to this passage along with the other three as "excerpts … *taken* [italic added] from two ELT textbooks" (p. 213). They do not, however, tell their readers that they have *modified* the "excerpts" as well. While the passage they have included in their C-Test consists of only 80 words, it comprises 181 in Alexander (1967, p. 217). In addition to shortening the text, they changed the constituting words of the passage. For example, the sentence "He *was obviously very* nervous …" is changed to "He *seemed extremely* nervous …" for no apparent reason. This means that conventional C-Test designers like Babaii and Ansari not only employ language *teaching* materials for *testing* purposes but also impose their own interpretations on what they choose. (The original passage developed by Alexander and the text simplified by Babaii and Ansari are given in Appendix.)

***Content Representation and C-Tests***

In contrast to the macro structural approach of schema theory which falls short of providing any objective and measurable unit of "special knowledge" (e.g., Grabe, 2002), its microstructural

perspective considers any word/phrase comprising an authentic text as a schema whose comprehension on the part of its readers depends on its meaning in relation to syntactic, semantic and discoursal relationship it holds with other schemata. [Interested readers are referred to Khodadady (2012) for more details.] Since the comprehension of each and all schemata comprising a given text determines their comprehension and thus behave as its main unit, they must be employed as the *best* and *only* units to develop test items on.

Grotjahn (2012a) seemed to be following the microstructural approach of schema theory because he believed that choosing one authentic text to develop AC-Test would under represent content in terms of *lexis*. He wrote

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 (p. 21).

Not only the first criticism dealing with the *alleged* content underrepresentation of AC-Test developed by K & H but also the second prescriptive criticism regarding the assumed "(severe) bias and unfairness" of the test are subjective because the data support the opposite as shown in Table 2. If we accept Richards, Platt and Platt's (1992) definition of lexis as "the vocabulary of a language in contrast to its grammar" (p. 213), it is best represented by semantic schemata, i.e., adjectives, adverbs, nouns and verbs. As can be seen, Dugatkin (2005) employed 866 semantic schemata to write the single authentic text consisting of three parts. K & H reproduced its introduction section to develop their AC-Test.

***Table 2: The frequency of semantic, syntactic and parasyntactic tokens comprising texts***

| Schema domains | Klein-Braley's four texts | | K&H's single text | | Dugatkin's (2005) text | |
|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| Semantic | 174 | 51.3 | 271 | 46.2 | 866 | 50.5 |
| Syntactic | 148 | 43.7 | 222 | 37.8 | 638 | 37.2 |
| Parasyntactic | 17 | 5.0 | 94 | 16.0 | 211 | 12.3 |
| Total | 339 | 100.0 | 587 | 100.0 | 1715 | 100.0 |

As it can also be seen in Table 2, the number of semantic schemata comprising the single AC-Test developed by K&H, i.e., 271, is greater than that of Klein-Braley's (1997), i.e., 174. The Chi-Square analysis of frequency showed that the difference in the number is statistically significant ($x^2$ = 21.144, $p$ <.001) and thus the AC-Test represents content/lexis significantly better than C-Tests. If we take content representation as a measure of fairness, then AC-Tests are fairer than the conventional C-Tests developed on short and modified texts. It is counter argued in this paper that these are the conventional C-Tests which are biased because their designers manipulate the texts syntactically, semantically and discousely to develop, pilot and select well functioning items. In other words, conventional C-Test writers *write* and/or *modify* a number of short texts dealing with general topics and pilot them in order to get their own desired response.

***Internal Validity of Conventional C-Tests and AC-Tests***

In addition to the statistically significant and higher representation of content, AC-Tests are superior to conventional C-Tests in terms of their internal validity. In order to support the

superiority, K&H followed scholars such as Baker (1989) and specified two indices which must be used *together* in order to establish it *statistically*. They announced, "for determining the discrimination power of items point biserial correlations ($r_{pbi}$) between the total test score and individual items were calculated and coefficient of 0.25 and higher were used along with acceptable *p*-values as indices of well functioning items" (p. 34).

Grotjahn (2012a), however, called the results reported by K&H "surprising in several ways" (p. 22). As one of the *allegedly* surprising results, he questions "the small range of item difficulties in the standard C-Test" of K & H which contradicts his "own extensive data sets and also to the data reported by Klein-Braley (1996), Jafarpur (1999) or Babaii and Ansari (2001)". After focusing on the reported range, i.e., .37 to .73, he wonders "whether the difficulty values in the standard C-Test are correctly calculated." Table 3, presents the frequency of correct responses given to each item on the conventional C-Tests. They were divided by the number of participants, i.e., 135, to get the p-values (PVs). As can be seen, the PVs are correctly calculated and reported. (The CRs were not given by K & H because they are conventionally considered redundant. Conventional C-Test designers like Klein-Braley (1997) do not, for example, provide their readers with their mean $r_{pbi}$ let alone the $r_{pbi}$ of each item.)

***Table 3: Correct responses (CRs) given to each item on the C-Test and their p-values (PV)***

| Item | CR | PV | Item | CR | PV | Item | CR | PV | Item | CR | PV | Item | CR | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 99 | .73 | 21 | 75 | .56 | 41 | 67 | .50 | 61 | 52 | .39 | 81 | 72 | .53 |
| 2 | 91 | .67 | 22 | 62 | .46 | 42 | 83 | .61 | 62 | 62 | .46 | 82 | 70 | .52 |
| 3 | 76 | .56 | 23 | 75 | .56 | 43 | 65 | .48 | 63 | 67 | .50 | 83 | 88 | .65 |
| 4 | 74 | .55 | 24 | 80 | .59 | 44 | 79 | .59 | 64 | 90 | .67 | 84 | 61 | .45 |
| 5 | 74 | .55 | 25 | 65 | .48 | 45 | 77 | .57 | 65 | 59 | .44 | 85 | 72 | .53 |
| 6 | 78 | .58 | 26 | 76 | .56 | 46 | 56 | .41 | 66 | 62 | .46 | 86 | 88 | .65 |
| 7 | 99 | .73 | 27 | 58 | .43 | 47 | 63 | .47 | 67 | 68 | .50 | 87 | 76 | .56 |
| 8 | 89 | .66 | 28 | 71 | .53 | 48 | 77 | .57 | 68 | 56 | .41 | 88 | 64 | .47 |
| 9 | 65 | .48 | 29 | 83 | .61 | 49 | 78 | .58 | 69 | 80 | .59 | 89 | 68 | .50 |
| 10 | 90 | .67 | 30 | 61 | .45 | 50 | 71 | .53 | 70 | 63 | .47 | 90 | 50 | .37 |
| 11 | 96 | .71 | 31 | 70 | .52 | 51 | 66 | .49 | 71 | 77 | .57 | 91 | 70 | .52 |
| 12 | 75 | .56 | 32 | 67 | .50 | 52 | 83 | .61 | 72 | 89 | .66 | 92 | 57 | .42 |
| 13 | 79 | .59 | 33 | 75 | .56 | 53 | 80 | .59 | 73 | 61 | .45 | 93 | 63 | .47 |
| 14 | 75 | .56 | 34 | 65 | .48 | 54 | 75 | .56 | 74 | 83 | .61 | 94 | 84 | .62 |
| 15 | 61 | .45 | 35 | 53 | .39 | 55 | 63 | .47 | 75 | 64 | .47 | 95 | 68 | .50 |
| 16 | 87 | .64 | 36 | 80 | .59 | 56 | 81 | .60 | 76 | 84 | .62 | 96 | 75 | .56 |
| 17 | 68 | .50 | 37 | 58 | .43 | 57 | 74 | .55 | 77 | 78 | .58 | 97 | 85 | .63 |
| 18 | 92 | .68 | 38 | 73 | .54 | 58 | 81 | .60 | 78 | 69 | .51 | 98 | 78 | .58 |
| 19 | 64 | .47 | 39 | 89 | .66 | 59 | 57 | .42 | 79 | 57 | .42 | 99 | 77 | .57 |
| 20 | 80 | .59 | 40 | 61 | .45 | 60 | 79 | .59 | 80 | 60 | .44 | | | |

Grotjahn (2012a) also stated that K & H should not have generalized from the results obtained on the administration of a single AC-Test to all AC-Tests and their generalization is, therefore, another surprising result which must be treated as a flaw. This statement is questionable at best because all research projects are conducted to generalize their findings otherwise there would be no use for their publication. Klein-Braley (1997), for example, made similar generalizations about testing procedures other than C-Tests. After administering the conventional C-Test employed by K & H along with the tests specified in Table 1 she declared that "the C-Test shows superior performance over the other test procedures in the categories difficulty level, reliability, validity, [and] factorial validity" (p. 71).

Furthermore, Grotjahn (2012a) claimed that "the reported number of well-functioning items in the AC-Test is not correct (at least according to Table 4)" and based on this claim he concluded that "it appears that K & H have taken into account only the values for the discrimination index" (p. 23). Both the claim and conclusion are unfounded because the 97 well functioning items, i.e., 3, 5, 10, 11, 14, 17, 18, 19, 21, 24, 26, 28, 29, 30, 31, 35, 37, 40, 42, 45, 46, 47, 50, 51, 52, 55, 56, 57, 59, 60, 62, 63, 65, 66, 68, 69, 70, 73, 77, 81, 82, 86, 89, 90, 92, 95, 100, 101, 104, 105, 106, 107, 108, 110, 111, 112, 113, 115, 118, 120, 122, 124, 125, 126, 130, 132, 133, 135, 137, 138, 139, 140, 141, 142, 143, 145, 146, 147, 148, 149, 150, 151, 152, 153, 155, 157, 158, 160, 161, 163, 165, 166, 170, 173, 174, 179, and 180, all have acceptable difficulty *and* discrimination indices, i.e., IFs between .25 and .75 and IDs equal to or higher than.25, which can be checked in Table 4.

And finally, Grotjahn (2012a) expressed surprises other than those brought up in previous paragraphs. Since they are all based on the mere assumptions made on the indices employed to determine item functioning, they will not be addressed. (For example, without checking the 97 well functioning items given in Table 4 and enumerated in paragraph above, he announces that "It *appears* [italics added] that K & H have taken into account only the values for the discrimination index" and then based on this totally ***subjective assumption*** he surprisingly subtracts 12 items from 97 claiming that they "are acceptable with regard to discrimination but not in terms of difficulty.") Aside from unfounded assumptions he bring up a peculiar objection regarding the standard deviations obtained on conventional C-Tests and AC-Tests which needs to be addressed separately.

***Comparing Conventional C-Tests with AC-Test: Standard Deviations***

In order to reject K & H's adoption of standard deviations (SDs) as indices of comparison between standard C-Test and AC-Tests, Grotjahn (2012a) argued that SDs "clearly depend on the range of the scale and in comparing standard deviations, one has to take this fact into account" (p. 23). Based on this argument he concludes, "therefore the conclusion that the AC-Test distinguishes best among the test takers, because it has the highest standard deviation, is not sufficiently substantiated …". This conclusion stands in sharp contrast to its interpretation by authorities such as Thornkdike and Hagan (1977) who declared that the SD "is a measure of variability that goes with the arithmetic mean. It is useful in the field of tests and measurements primarily as providing a *standard* [italics added] unit of measure having comparable meaning from one test to another" (p. 46).

Grotjahn (2012a) brought up the range of a scale in order to reject K & H's statement that standard deviations are "standardized by their very nature" (p. 35). He seems to have forgotten the fact that SDs are based on the arithmetic mean which derives their strength from normal

distribution and for this very reason SDs provide the best and simplest index to compare two measures such as conventional C-Tests and AC-Tests. According to Thorndike (2005),

This unvarying relationship of the standard deviation unit to the arrangement of scores in the normal distribution gives the standard deviation a type of *standard* [italic added] meaning as a unit of score. It becomes a yardstick in terms of which groups may be compared or the status of a given individual on different traits expressed. For example, if John's score in reading is 1SD above the mean and his score in mathematics is 2 SDs above the mean, then his performance in mathematics is better than his performance in reading (p. 49).

The results of language proficiency tests are used to reach educational decisions. As such they play a significant role in test takers' lives. According to the results obtained by K&H, while conventional C-Tests did not differentiate among many test takers because they obtained the same score, the AC-Test did accomplish the task because of its constituting items and higher magnitude of SD as shown in Table 4. As can be seen, seven test takers have, for example, scored 58 out of 99 on the conventional C-Test and their Z scores are all .40, indicating that they are of the same level of language proficiency. However, while only test takers four and five have obtained the same score on the AC-Test, i.e., 111 out of 180, test taker one's Z-score on AC-Test, i.e., 1.5, is over three times higher than his Z-score on the standard C-Test, i.e., .40, indicating that the former provides a much better measure of his proficiency as those of test takers 2, 3, 4, and 5 do. These differences are all reflected in the SDs of conventional C-Test and AC-Test, i.e., 11.358 and 21.589, respectively (see K & H's Table 2 on page 5 and 6).

***Table 4: The scores of seven test takers on the standard C-Test and AC-Test***

| Test taker | Standard C-Test | | AC-Test | |
|---|---|---|---|---|
| | Raw score | Z Score | Raw score | Z Score |
| 1 | 58 | 0.40108 | 130 | 1.55498 |
| 2 | 58 | 0.40108 | 113 | 0.76754 |
| 3 | 58 | 0.40108 | 112 | 0.72122 |
| 4 | 58 | 0.40108 | 111 | 0.6749 |
| 5 | 58 | 0.40108 | 111 | 0.6749 |
| 6 | 58 | 0.40108 | 103 | 0.30434 |
| 7 | 58 | 0.40108 | 98 | 0.07274 |

***Reliability Estimate of Conventional C-Test and AC-Test***

As a pioneering designer of conventional C-Tests Klein-Braley (1997) employed Cronbach's alpha to explore the reliability of her conventional C-Test on the basis of its individual gaps as reproduced in Table 5 below. K & H applied the same reliability estimate to their data. Surprisingly, however, Grotjahn (2012a) named and criticized Khodadady (2007) specifically for using the estimate and declared that "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" (p. 24).

***Table 5: Basic test statistics for grouped test procedures reported by Klein-Braley (1997, p. 68)***

| Test | Mean | SD | *P* | $rtt_{(ALPHA)}$ | $r_{(DELTA)}$ |
|---|---|---|---|---|---|
| ALLCLOZE | 10.72 | 4.79 | .27 | .66 | .65 |
| C-TEST | 51.15 | 16.71 | .52 | .85 | .70 |
| ALLMC | 29.90 | 5.44 | .60 | .51 | .70 |
| ALLCE | 22.86 | 9.28 | .46 | .75 | .65 |

As an alternative to Cronbach's alpha, Grotjahn (2012a) suggested several approaches whose descriptions run for five paragraphs! What the present researcher understands from the descriptions is that since there is a possibility of "local item dependence (LID) and correlated errors" (p. 24), conventional C-Test designers must employ different approaches whose application is as cumbersome as choosing a large number of texts to replace those which may not function as expected. These approaches are questionable as far as the present author is concerned simply because they are fundamentally utilized to overcome the problem created by *faulty items* and *associated errors*. Using statistics to overcome inherent problems with proficiency tests such as conventional C-Tests, it sounds both esthetically and logically unacceptable. The most feasible approach would be to use the alpha for the tests employed as K & H did or to discard their mal functioning items and to calculate their alpha by employing their well functioning items as Khodadady (2012) did.

## CONCLUSION

This study analyzed conventional C-Tests in terms of the texts chosen by their designers and contended that they are not based on reduced redundancy (RR) because their mutilation of words is systematic rather than based on probability. Neither is the selection of several texts justified in RR because it has nothing to do with "special knowledge" of a given unauthentic text as assumed by Klein-Braley (1997). The conventional C-Tests were developed originally to overcome the shortcomings of cloze tests as integrative/top down measures of language proficiency. They did, however, create shortcomings of their own when their designers adopted a prescriptive approach in their development.

Developing conventional C-Tests on several short and modified texts does not necessarily render them "genuine" as Grotjahn (2012a, p. 21) claimed it to do. Neither do conventional C-Tests measure "special knowledge**s**" because they do not address schemata as macro structures. They are, instead, developed on single words and should therefore be viewed as offshoots of micro structural approach of schema-theory. It is, therefore, suggested that instead of choosing several short texts and modifying them to serve testing purposes, normal/ authentic texts written for being read by literate English users be selected to write theoretically strong and empirically superior C-Test as K & H did.

## REFERENCES

Alexander, L. G. (1967). *Practice and progress: An integrated course for pre-intermediate students*. London: Longman.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: OUP
Baghaei, P. (2008). The effects of the rhetorical organization of texts on the C-Test construct: A Rasch modelling study. *Melbourne Papers in Language and Testing*, 13(2): 32-51.
Baker, D. (1989). *Language testing: a critical survey and practical guide*. London: Edward Arnold.
Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.
Davies, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell.
Dugatkin, L. (May 07, 2005). Why don't we just kiss and make up? *New Scientist* 2498, p. 35. Retrieved May23, 2005, from http://www.newscientist.com/channel/life/mg18624981.300.
Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23(3): 290-325.
Grabe, W. (2002). Dilemmas for the development of second language reading abilities. In J. C., Richards & Renandya, W. A. (Eds.). *Methodology in language teaching: An anthology of current practice* (pp. 276-286). Cambridge: CUP.
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. (2012a). Theoretical Misconceptions and Misuse of Statistics: A Critique of Khodadady and Hashemi (2011) and Some General Remarks on Cronbach's Alpha. *Iranian Journal of Language Testing*, 2(1), 20-27.
Grotjahn, R. (Ed.). (2012b). Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends. Frankfurt am Main: Lang (to appear).
Grotjahn, R. (2012a). Theoretical Misconceptions and Misuse of Statistics: A Critique of Khodadady and Hashemi (2011) and Some General Remarks on Cronbach's Alpha. *Iranian Journal of Language Testing*, 2(1), 20-27.
Heaton, J. B. (1988). *Writing English language tests* (new edition). Essex: Longman.
Jafarpur, A. (1999). Can the C-test be improved with classical item analysis? *System*, 27(1): 79-89.
Kelly, R. (1978). *On the construct validation of comprehension tests: an exercise in applied linguistics*. Unpublished PhD thesis, University of Queensland.
Khodadady, E. (1997). *Schemata theory and multiple choice item tests measuring reading comprehension*. Unpublished PhD thesis, the University of Western Australia.
Khodadady, E. (1999). *Multiple-choice items in testing: Practice and theory*. Tehran: Rahnama.
Khodadady, E. (2007). C-Tests: Method specific measures of language proficiency. *Iranian Journal of Applied Linguistics*, 10(2): 1-26.
Khodadady, E. (2012). Validity and tests developed on reduced redundancy, language components and schema theory. *Theory and Practice in Language Studies*, 2(3): 585-595. doi:10.4304/tpls.2.3.585-595.
Khodadady, E., & Hashemi, M. (2011). Validity and C-Tests: The role of text authenticity. *Iranian Journal of Language Testing*, 1(1): 30-41.
Khodadady, E., & Herriman, M. (2000). Schemata theory and selected response item tests: from theory to practice. In A. J. Kunnan (Ed.), *Fairness and validation on language assessment* (pp. 201-222). Cambridge: CUP.

Klein-Braley, C. (1985b). Advance prediction of test difficulty. In C. Klein-Braley, & Raatz, U. (eds.). Fremdsprachen und Hochschule 13/14: Themat-ischer Teil: C-Tests in der Praxis, Bochum: AKS, 23-41.

Klein-Braley, C. (1994). *Language testing with the C-Test. A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty*. Higher doctoral thesis (Habilitationsschrift), University of Duisburg.

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

Madsen, H., S. (1983). *Techniques in testing*. Oxford: Oxford University Press.

Richards, J. C., Platt, J., & Platt, H. (1992). *Longman dictionary of language, teaching and applied linguistics* (3rd ed.). Essex: Longman.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt am Main: Peter Lang.

Spolsky, B. (1973). What does it mean to know a language; or how do you get somebody to perform his competence? In J. W. Oller J. and J. R. Richards (Eds.). *Focus on the learner* (pp.164-76). Rowley, MA: Newbury House.

Spolsky, B. (2001) Closing the cloze. In H. Pürschel & U. Raatz (eds.) *Tests and Translation. Papers in memory of Christine Klein-Braley*. Bochum: AKS-Verlag.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson.

Thorndike, R., L., & Hagan, E. P. (1977). *Measurement and evaluation in psychology and education* (4th ed.). New Delhi: Wiley Eastern.

Yule, G. (2006). *The study of language* (3rd ed.). Cambridge: CUP.

## Appendix

Alexander's (1967) original text and the text upon which Babaii and Ansari (2001) developed their first C-Test

**A slip of the Tongue**

People will do anything to see a free show- even if it is a bad one. When the news got round that a variety show would be presented at our local cinema by the P. and U. Bird Seed company, we all rushed to see it. We had to queue for hours to get in and there must have been several hundred people present just before the show began. Unfortunately, the show was one of the dullest we have ever seen. Those who failed to get in need not have felt disappointed as many of the artistes who should have appeared did not come. The only funny things we hear that evening came from the advertiser at the beginning of the programme. He was obviously very nervous and for some minutes stood awkwardly before the microphone. As soon as he opened his mouth, everyone burst out laughing. We all know what the poor man should have said, but what he actually said was: 'This is the Poo and Ee Seed Bird Company. Good ladies, evening and gentlemen!' (Alexander, 1967, p. 217)

**A Slip of the Tongue**

On a variety show presented by P. and U. Bird Seed Company, a funny thing happened. It came from the advertiser at the beginning of the program. He seemed extremely nervous and for some minutes stood awkwardly before the microphone. As soon as he opened his mouth, everyone burst out laughing. We all knew what the poor man *should* have said, but what he *actually* said was: ″This is the Poo and Ee Seed Bird Company. Good ladies, evening and gentlemen!″ (Babaii & Ansari, 2001, p. 217)