The effects of the rhetorical organization of texts on the C-Test construct: A Rasch modelling study

Purya Baghaei

Islamic Azad University, Mashad, Iran

Abstract

In this study the effects of the rhetorical organization of C-Texts on the construct validity of C-Tests is investigated. Four passages with different rhetorical organizations were converted into C-Tests. The four rhetorical organizations were collection of descriptions, causation, comparison and problem/solution. The four C-Test passages were given to 104 subjects. The results of Rasch analysis of the data, studies of fit statistics, principal component analysis of residuals, and Smith's (2002) t-test show that some text-types require totally different strategies than other text-types. The results are in line with findings in text linguistics, schema theory and research in reading in first and second language.

1. Introduction

In the late 1970s and early 1980s when the cloze test had become an established test of overall language proficiency it came under severe attack by Alderson (1978, 1979, 1980, 1983) and Klein-Braley (1981). Alderson (1983) studied the effects of variations in deletion frequency and passage difficulty as determined by readability formulae and subjective ratings on psychometric characteristics of cloze tests. He demonstrated that although deletion rate affects test difficulty this is not predictable. Cloze tests with a high rate of deletion (n=6) constructed on easy and medium difficulty texts were less difficult than those based on the same texts with a lower deletion rate (n=12). He also concluded that the adjacent context around a gap when increased to more than five words did not help subjects to solve an item, so cloze tests he claimed are primarily tests of lower order skills (1979). He also found that the correlations of cloze tests constructed with different deletion frequencies with an external criterion - the ELBA (Ingram, 1964), changed considerably. This test, which was formerly used by English universities to examine their foreign students, had seven sections: (1) Sound Recognition, (2) Intonation, (3) Stress, (4) General Listening Comprehension, (5) Grammar, (6) Vocabulary, and (7) Reading comprehension (Alderson, 1983). Alderson interpreted this finding as the change of validity of the cloze (1983).

Melbourne Papers in Language Testing 2008 Volume 13.2 pp. 32-51. The Language Testing Research Centre, The University of Melbourne.

Klein-Braley's (1981) study also confirmed Alderson's findings. She addressed the automatic validity and reliability of cloze tests. She developed different cloze tests on the basis of nine passages by changing the deletion rate and the point of onset of deletion. She hypothesized that if all cloze tests are automatically valid and reliable, all these cloze tests should be equivalent. However, the results of her study showed that test difficulty depended on the text type, deletion rate and point of onset of deletion. She also compared the internal consistencies of different cloze tests and indicated that there were large differences in their internal consistencies. Hence, cloze tests are not parallel. She also noticed that the distribution of word classes namely, content words and function words affected by deletion in cloze tests, is not always representative of the distribution of the word classes in the intact passage. Factor analyses of her cloze tests and DELTA (an English placement test used at the University of Duisburg) revealed that some cloze tests loaded on the general factor while some did not. She also corroborated Alderson's claim that cloze tests are not tests of higher order skills. She correlated the items in cloze tests with each other and found out that there was no relationship between the distance of the items which were being correlated and the size of the correlations.

These considerations led Klein-Braley and Raatz (1982) to propose a new testing procedure based on the tenets of the cloze test but allegedly without its deficiencies. They called this new test type the C-Test, which is a variation of the cloze test and in fact the letter C stands for cloze to call to mind the relationship between the two tests. In this type of test the deletion rate is two, however instead of deleting the entire words the second half of words is deleted. The point of onset of deletion is also specified by the proponents: the second word in the second sentence. The advantages of C-Tests over cloze tests as mentioned by its advocates are that a C-Test battery is comprised of 4-6 short passages so as to eliminate text specificity and test bias associated with the cloze. Each passage has 20-25 gaps or items and the whole battery has 100 items so while being short it has many items using different texts. Advocates of the C-Test claim, and have empirically demonstrated, that adult educated native speakers usually obtain perfect scores on C-Tests, which is not the case with the cloze, at least when the exact word method of scoring is utilized. Furthermore, in C-Tests only exact-word scoring is possible.

However, we should bear in mind that these characteristics are not automatic as Raatz and Klein-Braley (1985) in their guidelines for constructing C-Tests implicitly and Grotjahn (1987) explicitly mention. 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. Besides, some texts do not correlate satisfactorily with the total test score and consequently should be discarded. Grotjahn (1987) also warns that texts with 20 deletions i.e. items, do not measure ‘macro-level textual constraints’ as the results of his think-aloud data analysis show. In sum, the C-Test was proposed to improve on the drawbacks of the cloze test. Spolsky (2001) criticizes the cloze and praises the C-Test:

> *By omitting words, which are linguistic 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 of the reduction of redundancy and so a purer example of an integrative rather than a discrete item test (p. 7).*

2. Text-level processing and the C-Test

There have been some assertions in the literature that C-Tests, due to their specific format, engage test-takers in micro-level processing and they are not involved in macro-level textual activities. By these two types of processing researchers mainly mean the amount of context that test-takers take into account when solving the gaps. The larger the amount of context on either side of a gap the test-taker processes (reads) when attempting an item the more the individual is involved in higher-order skills.

Several studies have been done to test whether C-Tests engage test-takers in text-level processing or not. Stemmer (1991/92) probed into this issue by means of think-aloud verbal data. She performed propositional analyses of three intact passages on which she constructed her C-Tests for the study. She defines proposition as the knowledge we have about facts which are stored in the memory in the form of some semantic units. A statement such as “John likes chocolates” is considered a proposition, or in other words, a proposition is comprised of a predicate and an argument. She identified the propositions in three texts and then converted them into C-Test passages. Stemmer’s main interest was to determine the frequency of crossing these propositions i.e. reading or taking into account contexts larger than a proposition, by individual test-takers. She found that only 12% of strategies that lead to solving and checking the suitability of a solution involve crossing propositional boundaries. And when subjects read part of a text to come up with a solution, only in 28% did they cross propositions. She concludes that C-Tests do not involve test-takers in high-level comprehension.

Grotjahn and Tönshoff (1992) used a different design to check whether comprehension takes place when individuals solve C-Test items. They gave C-Test passages which were 50-61 items long to subjects. The subjects were required to solve the test items, hand them in, and then jot down what they remembered about the content of the passage. Then they were given the intact passage to translate it into their L1 so as to have a criterion to judge whether subject's inability to recall the content of the passage was because C-Tests do not allow text comprehension or it was because of the difficulty of the text for the target group. The researchers found high to medium correlations between C-test scores and the number of correct idea units recalled by test-takers and their translation scores. However, as the C-Test passages used in this study were twice as long as the standard C-Test passages we should be careful about generalizing the results to C-Test passages with only 20-25 items as researchers themselves also mention. Their final conclusion is that the C-Test does tap reading comprehension but not in its standard 20-25 item length.

Grotjahn (1996) studied the effects of text scrambling in C-Test passages on the difficulty of the items. He concluded that disrupting the connectivity of texts, which makes them more demanding to process should increase C-Tests difficulty. But the effect was only observed in difficult and longer passages. He concluded that C-Tests of standard length measure on the micro context level and if C-Tests are supposed to measure text-level macro skills longer passages are required.

However, Grotjahn (2002) replicated his 1996 study with short C-Test passages and concluded that C-Tests in their standard length can also measure higher-level skills while the effect of scrambling proved to be stronger for more difficult passages as well as for more proficient test-takers.

3. The Problem

Views on whether C-Tests involve test-takers in higher-order skills are far from unanimous. A major contribution to this aspect of C-Test research is that of Sigott (2004), where he enters text and test-taker attributes into the arena, implying that discussions of the C-Test construct and its potential for triggering higher-order skills without consideration of text and test-takers attributes are unwarranted. He introduces the concept of the *fluid construct phenomenon (FCP)*; he basically argues that the C-Test construct changes as a function of text difficulty and test-taker ability. He empirically demonstrates that the construct of C-Test changes for test-takers of differing ability levels. He shows that high-proficiency students manage to solve the C-Test items based on small portions of the text on either sides of each blank, hence, less text-level or higher order processing for these test-takers. While low-proficiency students, in order to solve the items need to read larger portions of texts and combine information from different parts of the text, hence, more higher-order or text-level processing for these test-takers. He produces ample evidence that supports his claim as regards test-taker ability but none regarding text difficulty, although, in the formulation of the

FCP he mentions both test-taker ability and text difficulty. That is, he claims, the difficulty of the texts which are converted into C-Test is, in addition to test-taker ability, a determinant of what the C-Test measures.

The present study focuses on another aspect of text and its effect on C-Test construct, i.e., the rhetorical organization of texts. The research question addressed in this study is:

Does the rhetorical organization of texts which are converted into C-Tests affect the C-Test construct?

In other words, is the C-Test construct a function of the rhetorical organization of the texts? The approach adopted here is totally empirical. Techniques available within Rasch modelling are employed to answer the question. These techniques which mainly focus on ascertaining unidimensionality, a major Rasch measurement principle, best suit to answer this question.

4. Procedures

As mentioned above the effect of the rhetorical organization of texts which are mutilated to be used as C-Tests on the extent of text-level processing triggered is the focus of this study. The effect of four text types are studied here, namely, collection of descriptions, causation, comparison and problem/solution. This is Meyer's (1975, 1979) classification of expository discourse which she empirically shows affects reading comprehension in English as a native language. Carrell (1984, 1985) studies the effect of these rhetorical organizations on the reading comprehension of ESL students and finds out that text structure affects comprehension and recall of information.

Because of the local dependence of gaps (items) in C-Tests it is a common to consider each passage as a super-item. Many statistical operations including Cronbach's alpha reliability estimates, Rasch model and all IRT models assume local item independence in the data. Dependence of items on each other which is prevalent in cloze tests, C-Tests and reading comprehension tests when the questions are based on a single passage, inflates correlations among the items and results in spurious reliabilities and artificially small standard errors. Considering each passage an item and entering them into the analysis as independent single items solves the problem of local item dependence in C-Tests. In the context of C-Test, research passages are referred to as super-items and gaps are called micro-items.

4.1 Instruments

Four C-Test passages each based on a different rhetorical organization were constructed. In fact, the texts which were used by Carrell (1984) in her study as reading comprehension tests were converted into C-Test passages (with permission) by deleting the second half of every second word, leaving the first and last sentences intact.

After doing so, in the first passage, which was a collection of descriptions text, there were 50 micro-items (blanks), in the second passage which was a causation type there were 41 micro-items, in the third passage, a comparison text, 63 and in the fourth passage, a problem/solution text, there were 51 micro-items. The texts were too long to produce canonical C-Test passages with 20-25 blanks. In order to avoid changing the rhetorical structure of the texts, they were not shortened and were used in their full length.

Alongside the C-Test battery, a reading comprehension test was also administered. The test consisted of two parts. In the first part there was a long passage followed by eight multiple choice questions. In the second part there was another passage with six missing sentences. The missing sentences were on another page with an extra sentence. The students were required to find out where in the text the missing sentences fit. The places in the text where the sentences had been removed were numbered. The multiple choice items were named MC1-MC8 and the sentence insertion items were named SI1-SI6. The reading comprehension test was intended to be a test of text-level processing. Text-level processing was defined as reading or processing beyond sentence, i.e., crossing sentence level boundaries in producing reply to a question. Each reading comprehension question was supposed to trigger text-level skills. This means that for answering the reading questions, students needed to read and process at least two sentences.

The subjects were 104 Iranian undergraduate students of English at different years of their studies. The whole test battery containing the four C-Test passages, the 8 multiple-choice reading comprehension questions and the sentence insertion task with 6 deleted sentences were given to the subjects in the same order within a hundred- minute teaching session.

In order to confirm that the reading test is actually a text-level test a group of university English instructors were asked to fill in a questionnaire and answer whether for replying each multiple choice question students need to read beyond a sentence and combine information from at least two sentences. The majority of the instructors agreed that all but one multiple choice reading comprehension question elicit text-level strategies. This item which was considered a vocabulary item was eliminated.

Considering the definition of text-level processing given above, the second part of the reading comprehension test, i.e., the sentence insertion (SI) task could well be called a text-level processing test. Finding the right places of the removed sentences entails reading and understanding the sentences and paragraphs that precede and follow the gaps. This, no doubt, happens when one wants to answer such questions. All in all, the reading comprehension test is considered a text-level processing test.

Cronbach alpha reliability of the whole battery was 0.75; that of the C-Test passages 0.91 and that of the reading comprehension items only 0.57. It should be borne in mind that the low reliability of the reading test can partially be an artifact of the small number of items. The reading test and the C-Test significantly correlated at 0.59 ($p< 0.01$; $n=104$).

5. Results

5.1 Rasch analysis

As was explained earlier, one of the assumptions of the Rasch model is local item independence. Since this assumption is violated in C-Tests, the passages which are independent of each other are entered into the analysis as single items. Therefore, here we have only four items in the analysis. For the analysis WINSTEPS Rasch Programme (Linacre, 2006) was used.

Because each item (passage) had its own rating scale structure (due to unequal number of blanks in each passage), Master's (1982) partial credit model as implemented in WINSTEPS was used for Rasch analysis. Partial credit model is a member of the family of Rasch models which handles polytomous items. Unlike dichotomous items where the replies of test-takers are scored as either totally correct (1) or totally incorrect (0), in polytomous items test-takers' replies can be scored on a number of levels, say between 0-5, depending on the quality of the reply provided. Master's Partial credit model handles such data.

As Table 1 shows, the collection of descriptions text (SUPERC1) grossly misfits. This is an indication of the lack of unidimensionality. All four fit indicators clearly show the misfit of collection of description. The acceptable range for mean square statistics according to McNamara (1996) is 0.70-1.30 and for ZSTD is -2 to+2. As you can see in the table the infit and outfit mean squares for this item is 1.81 and 1.83 respectively and the values for infit and outfit ZSTD's are 4.5 and 4.7 respectively; way above their acceptable values. SUPERC3 (comparison) is an overfitting item. Its small infit and outfit mean squares and negative ZSTD's show the predictability or redundancy of this item. Such items do not degrade measurement and are not threat to construct validity; they only do not add anything to our information about the construct (Linacre, 2006).

Table 1: Measure order and fit statistics for the C-Test passages

Entry Number	**Raw Score**	**Measure**	**S.E.**	**Infit MNSQ**	**Infit ZSTD**	**Outfit MNSQ**	**Outfit ZSTD**	**PTMEA Corr.**	**Item**
1	2769	.39	.03	1.81	4.5	1.83	4.7	.85	SUPERC1
2	2518	-.10	.03	.93	-.4	1.04	.3	.86	SUPERC2
4	3122	-.13	.03	.82	-1.2	.83	-1.2	.87	SUPERC4
3	4456	-.16	.02	.60	-2.8	.57	- 3.4	.85	SUPERC3
Mean	3216.3	.00	.03	1.04	.0	1.07	.1		
S.D.	747.2	.23	.00	.46	2.7	.47	2.9		

In raw score terms SUPERC2, i.e., the causation text is the most difficult item but in terms of Rasch measures it has a lower measure than the collection of descriptions text which is an indication of the misfit of the collection of descriptions. The misfit of the collection of description text is evident from its out of range mean square and ZSTD statistics. Collection of descriptions, according to Carrell (1984), is the least tightly organized discourse type and the other three are more highly structured. Carrell (1984, 1985), within the context of schema theory research, shows that the text structure interacts with the schemata of readers. According to Carrell (1984), each reader has a limited number of text organization schemata. When a reader embarks on reading a text she approaches the text with her knowledge of how texts are usually organized. The reader searches in her repertoire for the text organization that best accounts for the text at hand. The reader also activates the same text organization schemata during recall to retrieve information from memory. Meyer and Freedle (1984) and Carrell (1894) show that the recall of idea units from the more structured causation, comparison and problem solution text types is much easier than the recall of idea units from the loosely organized collection of description text type both for native English readers and for ESL readers. The reason, Carrell (1984) argues, is that in these text types

> *the superordinate structure gets rehearsed with each new piece of information that the reader processes and attempts to integrate with the main ideas of the text. … The causation, problem/solution, and comparison structures are more highly organized types of top level structures than the collection of descriptions top-level structure because of the particular relationships which hold between the top-level nodes in the former structures. There are no particular relationships holding between the top-level nodes in the latter structure…the causation, problem/solution, and comparison structures, with their more highly organized components, are expected to facilitate encoding, economy of storage in long-term memory, and subsequent retrieval processes (p.447).*

Passage 3, the comparison text has a very small mean square fit statistics which indicate overfit, i.e., too much of predictability, redundancy, or local dependence, though this cannot be considered a threat to the measurement. Causation and problem/solution texts have also mean square statistics smaller than one, though acceptable, which mean predictability and dependency. It is surprising that these three text types overfit while they are totally independent of each other. Carrell (1984) states that “Discourse organized with the causation, problem/solution, and comparison structures contains a number of overlapping issues viewed from different perspectives” (p.447). This could well be the reason why these three texts overfit, although they are locally independent items. The rhetorical organization of these three text types which affect processing strategies overlap. That is, the dependency is in processing strategies and not in the item content which results in overfit. So the reason for overfit should not necessarily be local dependency among the items; similarity of strategies used to process items can also result in overfit.

Lee (2004) lists possible sources of local item dependence (LID), such as “the sharing of a common passage, content, knowledge, item chaining, speededness, fatigue, practice effects, item or response format, and so forth. Muraki and Lee (2001) reported that the physical layout of the test booklet could also become a potential source of LID for some items in the test” (p. 76). Lee (2004) has suggested that even items which do not share a passage may also be locally dependent:

> *…a certain type of items could seem difficult at first, but become somewhat easier with practice on similar items, due to some cognitive tasks or attributes shared among all items of this type. It follows then that, when the similar item types are used for multiple passages, item types can create an additional LID factor due to their test method effect (Bachman, 1990) that can potentially be overlapping with the influence of a shared text … Hence, an intriguing question would be: Is it possible that a common cognitive attribute (linguistic or nonlinguistic) shared among items of the same type can make all items sharing that cognitive feature locally dependent due to common secondary factors unaccounted for by the IRT theta or the practice effects over similar items in a test, as pointed out by Rosenbaum (1988)? (p. 79f.).*

As the results of the present study show, the answer to this question raised by Lee is ‘yes’. The underfit of the collection of descriptions and overfit of the causation, comparison, and problem/solution C-Test texts are well in line with the findings of research in schema theory in reading in English in first and second language. As mentioned earlier, decoding and recall of information from a collection of description text is more difficult than from other three text-types because in collection of description texts there is no relationship or connections between the components of the text.

5.2 Principal component analysis of residuals

Principal component analysis of residuals is a technique to assess unidimensionality in the data. Residuals are the differences between the expected probabilities of correct replies to items (Rasch model) and real observations. The smaller the differences between real observations and model expectations, the better the model has accounted for the data, hence, better fit between the data and the model. In fact, the fit statistics considered above are computed on the basis of the residuals. Residuals form part of the data that the model has not explained, so we expect them to be uncorrelated and be randomly distributed. In a factor analysis of residuals we expect to find no structure. However, if we do find a structure it means that the data are not unidimensional and a subsidiary dimension has contaminated the data.

Principle component analysis of residuals for these data indicates that the test is not unidimensional. Table 2 summarizes the results of principal component analysis of residuals for these data. The first column in the table shows the sizes of the variances in eigenvalue units. The second column is the percentage sizes of the variances in log unit and the third column is the Rasch model expectations, i.e., how the situation would be if the data fit the Rasch model perfectly. The Rasch dimension explains 95.2% of the variance in the data. It is very close to the model expectation of 95.5%.

The unexplained variance in the data is 4.8%, this includes the Rasch-predicted randomness and any departures from Rasch criteria, e.g., multidimensionality (Linacre, 2006). The strength of this contrast is 1.6 in eigenvalue units which means that the subsidiary dimension has the strength of 1.6 items which is huge in the case of a four-item test.

Table 2: Table of standardized residual variance

	In Eigenvalue units	**In percentages**	**Modeled**
Total variance in observations	83.3	100.0%	100.0%
Variance explained by measures	79.8	95.2%	95.5%
Unexplained variance (total)	4.0	4.8%	4.5%
Unexplained variance - 1st contrast	1.6	1.9%	
Unexplained variance - 2nd contrast	1.2	1.5%	
Unexplained variance - 3rd contrast	1.1	1.3%	
Unexplained variance - 4th contrast	0.0	0.0%	
Unexplained variance - 5th contrast	0.0	0.0%	

Table 3 below shows the items that form contrast one or the unwanted secondary dimension. Lower case letters and capital letters indicate items with most opposed loadings on the first contrast. As Table 3 shows, the first contrast partitions the items into two clusters, A vs. a, b and c. Table 3 also identifies these clusters, SUPERC1 vs. SUPERC2, SUPERC3 and SUPERC4. Positive and negative loadings are arbitrary here and only show the direction. SUPERC2 and SUPERC3 which have the highest negative loadings on the contrast form one end and SUPERC1 which has the highest negative loading on the contrast form the other end. In tests where there are more items, the loadings in both direction get smaller as we go down the loading column. The two ends of the contrast are the first few items in each direction which have the highest loadings. We do not know which direction the contrast goes; this requires substantive content analysis of the items. However, we do know that one end is more the dimension that we intend to measure. "The convention is that the majority cluster sets the standard, i.e., the majority cluster is the intended Rasch dimension and the minority cluster is the secondary dimension" (Linacre, personal communication). Here we have one item in one cluster and three in the other. Moreover, SUPERC1 (collection descriptions) misfits. So, the cluster of SUPERC1 should be the cluster that opposes the Rasch dimension and the rest should define the latent trait.

Table 3: Contrast 1 table sorted by loading for the C-Test passages[*]

Contrast	Loading	Label	Item
1	.96	A	SUPERC1

Contrast	Loading	Label	Item
1	-.60	a	SUPERC3
1	-.56	b	SUPERC2
1	-.09	c	SUPERC4

In the following section the four C-Test passages and the two reading tasks are entered into a combined analysis. The reading tasks are treated as two polytomous item here. Table 4 displays that when the reading tasks are entered into the Rasch analysis as two polytomous items, the collection of descriptions text misfits and the rest fit although the third and fourth C-Test passages, i.e., the comparison text and the problem/solution text overfit. The misfit of collection of descriptions text (SUPERC1) shows that there is something systematically different about this text in comparison to the other

[*] There are two tables as Contrast 1 table and they should be considered together; one for negatively loading items and one for positively loading ones.

texts and the reading tasks. The overfit of the comparison text and the problem/solution text show the redundancy of these texts and the reproducibility of the information provided by these texts.

Table 4: Measure order and fit statistics for C-Tests and reading tasks

Entry Number	**Raw Score**	**Measure**	**S.E.**	**Infit MNSQ**	**Infit ZSTD**	**Outfit MNSQ**	**Outfit ZSTD**	**PTMEA Corr.**	**Item**
20	233	.78	.11	1.21	1.5	1.25	1.6	. 37	SITOTAL
15	2769	.14	.03	1.67	3.8	1.64	3.8	.85	SUPERC1
19	406	.05	.07	1.08	.6	1.04	.3	.69	MCRCTOTAL
16	2518	-.29	.03	.89	-.7	.95	-.3	.86	SUPERC2
18	3122	-.31	.03	.74	-1.7	.77	-1.6	.87	SUPERC4
17	4456	-.36	.02	.59	-2.9	.58	-3.4	. 84	SUPERC3
Mean	2250.7	.00	.05	1.03	.1	1.04	.1		
S.D.	1496.5	.39	.03	.35	2.2	.34	2.3		

The principal component analysis of standardized residuals when the two reading tasks are added, confirms multidimensionality as the first contrast has the strength of about two items (1.9) out of six. In fact Smith (2002) considers eigenvalues greater than 1.5 as representing the existence of a second dimension.

Table 5: Table of standardized residual variance

	In Eigenvalue units	**In percentages**	**Modeled**
Total variance in observations	86.4	100.0%	100.0%
Variance explained by measures	80.4	93.1%	93.3%
Unexplained variance (total)	6.0	6.9%	6.7%
Unexplained variance - 1st contrast	1.9	2.1%	
Unexplained variance - 2nd contrast	1.3	1.5%	
Unexplained variance - 3rd contrast	1.2	1.4%	
Unexplained variance - 4th contrast	0.9	1.1%	
Unexplained variance - 5th contrast	0.7	0.8%	

Table 6: Contrast 1 table sorted by loading for C-Test passages and reading tasks

Contrast	Loading	Label	Item
1	.78	A	SUPERC1
1	.58	B	MCRCTOTA
1	.42	C	SITOTAL

Contrast	Loading	Label	Item
1	-.69	a	SUPERC3
1	-.49	b	SUPERC2
1	-.16	c	SUPERC4

Table 6, the contrast one table for C-test passages plus reading tasks, shows that the collection of descriptions text and the reading tasks oppose the other texts. If we consider the two reading tasks as activities which trigger higher-order skills , as they were intended to be , then one is inclined to assert that collection of descriptions text triggers more text-level skills than causation, comparison and problem/solution texts because it cluster with the reading tasks. This also could be the reason why this text does not fit when the four C-Test passages are Rasch analysed together. In other words, the type of processing or reading strategies which are triggered by the causation, comparison and problem/solution are similar to each other and different from those which are triggered by the collection of descriptions text. As was mentioned above, these three rhetorical organizations have a number of overlapping issues (Carell, 1984).

5.3 The t-test approach

A very recent approach in detecting dimensionality is Smith's (2002) t-test approach. In the Rasch model, the ability estimates of persons should not depend on the subset of items that they happen to encounter, if the data fit the Rasch model. That is, if we divide the items in a test into two subsets and estimate persons' abilities separately once on the basis of Subset *X* and once on the basis of Subset *Z*, after a translation necessary to set an origin common to the two analyses to account for the change in the local origin of the two scales (Wright and Stone, 1979), we should get equivalent person ability estimates within measurement error. In this approach, t-tests are conducted to check if the person ability measures from the two analyses are equivalent. The t-tests to check whether person measures have statistically remained invariant are computed according to the following formula (Wright and Stone, 1979; Smith 2002):

$$t = B_x - B_z / \sqrt{SE_x^2 + SE_z^2}$$

Where B_x is the person measure from subset X, B_z is the person measure from subset Z (after being brought onto the same scale), SE_x and SE_z are the standard errors of person measures from the two analyses. The t-test is conducted for every person in the analysis. A statistically significant t-test indicates that the person's ability measure depends on the subset that s/he has taken, hence, there is multidimensionality which is a violation of the Rasch model assumptions. One should bear in mind that unidimensionality is not an absolute matter rather it is a matter of degree. The concern of the instrument designer should be whether the defined variable is unidimensional enough so that it does not affect person measures (Smith, 2002). Statistical insignificance of these independent t-tests is an indication of unidimensionality. In other words, the ability estimates of persons do not depend on the subset of items. In order to be sure that the measure is "usefully" unidimensional we can construct the familiar 95% quality control lines around the identity line with slope of unity in the plot of person measures. We expect that 95% of the persons fall within the control lines.

Here this t-test approach which, according to Tennant and Pallant (2006) is the most robust method in identifying multidimensionality is used to check the unidimensionality of the C-Test. As was shown by the principal component analysis of residuals, the collection of descriptions text sharply contrasted against causation and comparison texts. Loadings of above 0.30 were considered as significant (Tennant and Pallant, 2006). These items have loadings of above 0.30 in either direction on the first contrast. Therefore, the subsets of items to compare person measures against each other were made up of the collection of descriptions text (SUPERC1) against causation plus comparison texts (SUPERC2 and SUPERC3). Since the loading of the problem/solution text (SUPERC4) was only 0.09 it was not included in the analysis. So, for each person two ability measures were estimated; one on the basis of the collection of descriptions text and one on the basis of the causation and comparison texts. For each ability estimate there is an associated standard error. The ability estimates and their standard errors were used to compute 104 t-tests according to the equation above.

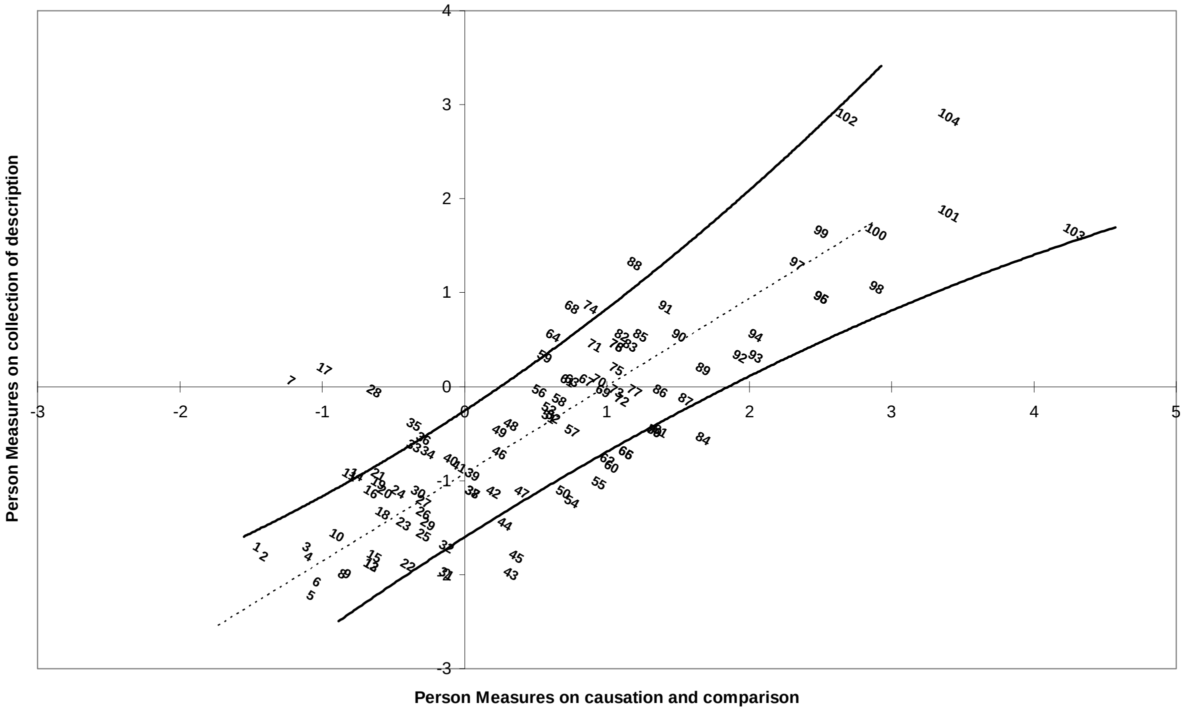
In case of significant differences in person measures on the basis of these two subsets, one can argue that the rhetorical organization of texts affect the strategies and skills triggered by the C-Tests which are constructed out of them, whatever these strategies and skills are. In the face of the evidence that we have from the two reading tasks we assume that these strategies are text-level and low-level skills since in the principal component analysis of residuals of the C-Tests and the reading tasks the collection of descriptions text and the reading tasks, which are supposed to be text-level tests, formed one cluster.

T-tests showed that none of the differences between person measures from the two subsets were statistically significant ($p < 0.05$). However, many of the ability measure differences were too large to be ignored just because of statistical insignificance. Thirty-four out of 104 differences are above 0.50 logits and 10 of them are above one logit. Half a logit difference in a person's measure can make a lot of difference in high-stakes assessments where there is a cut-off score to be reached for certain admission or selection purposes. Students' measures from the two item classes correlate at 0.78 which, though high, is very far from perfect. Therefore, if we are making norm-referenced decisions, the rank order of persons does change as a result of what section of the test is used.

The measures from the two subsets can be plotted on *x* and *y* axes in order to graphically display whether the two subsets of items have produced equivalent person ability measures. If the persons fall close to a diagonal line with a slope of one, then the two subsets have produced equivalent estimates, if they fall far from it then the estimates from the two analyses are not equivalent. Figure 1 shows the plot of measures from the subset of collection of descriptions vs. causation and comparison. The two parallel lines set the boundary for equivalent estimates. Persons who fall between these lines have had equivalent measures on the two subsets (within measurement error).

The cross plot of measures on collection of descriptions text against measures on causation and comparison texts with 95% control lines shows that many persons fall outside the control lines, in spite of insignificant t-tests. To be able to claim that the two subsets have produced equivalent measures, at least 95% of the persons should fall between the two lines.

Figure1: Cross plot of person measures from the two subsets



The principal component analysis of standardized residuals when the two reading tasks were entered into the analysis showed that the collection of descriptions text and the two reading tasks contrasted against other text types. Smith's t-test approach was used again to assess unidimensionality of the instrument. The collection of descriptions text and the reading tasks formed one subset and the other three text types, namely, causation, comparison and problem/solution formed the other subset. The WINSTEPS Rasch programme was used (Linacre, 2006) to obtain two measures for each person on the two subset of items, their difference, t-tests and their probabilities to check the statistical significance of the differences. The observed differences between the measures from the two subsets suggest that the two subsets measure different constructs and are two separate dimensions. Since here we know that the structures of the texts in the two subsets are different, one can aptly conclude that text structures or rhetorical organization of the texts has brought about the observed differences and therefore different text structures can lead to the measurement of different constructs in C-Tests.

Person measures on the two subsets were cross plotted against each other on *x* and *y* axes. The 95% quality control lines were drawn. Around 25 persons fell outside the control lines which indicate that for these persons the measures on the two subsets are not equivalent which is a sign of multidimensionality. In other words, the collection of descriptions text plus reading tasks is a different dimension from the combination of causation, comparison and problem/solution texts. This means that the C-Test battery comprised of the four text-types is not unidimensional and each text triggers different strategies.

6. Discussion

In this study an attempt was made to demonstrate whether the rhetorical organizations of the texts which are converted into C-Tests affect the constructs which are tested by them or not. A C-Test battery which contained four passages, each passage having a different rhetorical organization was used in this study. Rasch analysis of the data ascertained multidimensionality. Multidimensionality indicates that more than one construct is being tapped by the test. Since this C-Test battery was made up of texts with different rhetorical organizations the multidimensionality of the four C-Test passages suggests that different rhetorical structures trigger text-level skills to different degrees. Further investigations showed that in fact the collection of descriptions text-type is the problematic text-type that does not go with the other text-types. The other text-types which were used in this study, namely, causation, comparison, and problem/solution psychometrically agree and trigger the same skills. This is in line with the findings in text linguistics which asserts that causation, comparison and problem/solution discourse share many common features. As a result of their overlap, they tend to trigger the same strategies and consequently form one psychometric dimension. The collection of descriptions text-type is textually or structurally different from the other three text-types and requires different strategies to be processed and solved, therefore, it forms a different dimension.

Collection of description texts are difficult to process even as intact reading texts (Carrell, 1984), let alone when mutilated as C-Test passages. This could be the reason why it misfits. This finding is clear evidence that the processes involved in reading these unmutilated texts are the same when these texts are converted into C-Tests. This is an indication that C-Test taking is not a local puzzle solving task and is more or less an activity like reading unmutilated texts. Otherwise these four text-types would have behaved similarly in Rasch analysis after being converted into C-Tests. Hastings (2002) who used error analysis as a window to what goes on in the mind of the C-Test takers states the same idea:

> *…the processing that is required for successful C-test performance seems comparable to natural language processing in both depth and*

> *complexity, and may in fact have much in common with natural language performance (p.66).*

The empirical evidence provided here shows that the processes involved in reading the intact texts are transferred to the C-Test-taking context when these texts are converted into C-Tests. Therefore, one can argue that different rhetorical organizations trigger text-level skills to different degrees in C-Tests because the intact texts require different skills and processes to be decoded.

It seems that collection of descriptions text trigger more text-level skills than other text types. The main evidence that supports this argument is that in the principal component analysis of residuals this text type clusters with the two reading tasks (which were intended to be tests of text-level processing) against the other three text-types. A possible psycholinguistic reason for the phenomenon that the collection of description text-type triggers more text-level skills is suggested here: According to schema theory, reading involves the process of verifying the formal schemata of a text that accounts for it. A skilled reader has a limited number of such text processing schemata and approaches reading with the knowledge of how texts are formally organized. For each particular text the reader searches her repertoire to find the best text-processing schema for the text at hand (Carrell, 1984). Since collection of description is not a formal and recognized text organization, there is certainly no schema in the reader's repertoire for processing this type of text. Therefore, students mostly rely on inside-text resources for decoding. That is, it is bottom up. Collection of description schema does not exist in the mind. The only thing that the reader can resort to is the text itself. There is less to help outside the text for this kind of rhetorical organization. And as we know reading is an interaction between the text and our background knowledge and schemata. It is both a top-down and a bottom-up processing. Since in the case of this text-type there is less chance to activate schemata, all effort is made to use the text and the features of the text. Hence, more text-level skills are triggered. This, of course, is not a definite psycholinguistics justification for the empirical evidence observed here, and the reason why some texts tend to trigger more text-level processing remains an open question.

The present study demonstrates that text structure influences what is measured by a C-Test. This is only one aspect of the text which is considered here. There are other text characteristics which are also interesting to study. This phenomenon can be exploited to develop C-Tests for measuring specific constructs. One ripe avenue of research in this area is to investigate if the use of spoken texts or conversations in the construction of C-Tests can lead to the measurement of oral skills. Investigating the effects of the characteristics of discourse on C-Test performance can help us better understand the construct of the C-Test and direct us towards a better appreciation of the merits and limitations of this test.

7. References

Alderson, J.C. (1978) *A study of the cloze procedure with native and non-native speakers of English*. Unpublished Ph.D dissertation, University of Edinburgh.

Alderson, J.C. (1979a) The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13:2, 219-227.

Alderson, J.C. (1979b) The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading*, 2:2, 108-119.

Alderson, J.C. (1980) Native and non-native speaker performance on cloze tests. *Language Learning*, 30:1, 59-76.

Alderson, J.C. (1983) The cloze procedure and proficiency in English as a foreign language. In J.W.Jr. Oller, (ed.) *Issues in language testing research*. Rowley MA: Newbury House.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly*, 19:4, 727-752.

Carrell, P. L. (1984). The effects of rhetorical organization on ESL readers. *TESOL Quarterly*, 18:3, 441-469.

Grotjahn, R. (1996). 'Scrambled C-Tests': Untersuchungen zum Zusammenhang zwischen Lösungsgüte und sequentieller Textstructur. In Rüdiger Grotjahn (ed.), *Der C-Test. Theoretische Grundlagen und Praktische Anwendungen*. Vol.3. Bochum: Brockmeyer.

Grotjahn, R. & Tönshof, W. (1992). Textverständnis bei der C-Test Bearbeitung. Pilotstudien mit Französisch- und Italienischlernern. In Rüdiger Grotjahn (ed.), *Der C-Test. Theoretische Grundlagen und Praktische Anwendungen*. Vol.1. Bochum: Brockmeyer.

Grotjahn, R. (2002). 'Scrambled' C-Tests: eine Folgeuntersuchung. In Rüdiger Grotjahn (ed.), *Der C-Test. Theoretische Grundlagen und Praktische Anwendungen*. Vol.4. Bochum: AKS-Verlag.

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 measures: the validity and validation of language tests*. Bochum: Brockmeyer.

Hastings, A. J. (2002). Error analysis of an English C-Test: Evidence for integrated processing. In Rüdiger Grotjahn (ed.) *Der C-Test. Theoretische Grundlagen und Praktische Anwendungen*. Vol.4. Bochum: AKS-Verlag.

Ingram, E. (1964) *English language battery (ELBA)*. University of Edinburgh: Department of Linguistics.

Klein-Braley, C. (1981) *Empirical investigations of cloze tests*. Unpublished doctoral dissertation, University of Duisburg.

Klein-braley, C. & Raatz, U. (1982) Der C-Test: Ein neuer Ansatz zur Messung von allgemeiner Sprachbeherrschung. *AKS-Rundbrief* 4, 23-37.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, 21:1, 74-100.

Linacre, J. M. (2006). WINSTEPS Rasch programme. Version 3.63.0.

Linacre, J.M. (2006) *A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs*. Chicago, IL: winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47:2, 149-174.

Meyer, B. (1975). *The organization of prose and its effects on memory*. Amsterdam: North Holland Publishing Co.

Meyer, B. (1979). Organizational patterns in prose and their use in reading. In Michael L. Kamil, & Alden J. Moe (eds.) *Reading research: studies and applications*. Clemson, South Carolina: National Reading Conference.

Meyer, B. & Freedle, R. (1984). Effects of discourse type on recall. *American Educational Research Journal*, 21:1, 121-143.

Raatz, U. & Klein-Braley, C. (1985) How to develop a C-Test. *Fremdsprachen und Hochschule* 13/14, 20-22.

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

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3:2, 205-231.

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.

Stemmer, B. (1991). *What's on a C-Test taker's mind? Mental processes in C-Test taking*. Bochum: Brockmeyer.

Stemmer, B. (1992). An alternative approach to C-Test validation. In Rüdiger Grotjahn (ed.) *Der C-Test. Theoretische Grundlagen und Praktische Anwendungen*. Vol.1. Bochum: Brockmeyer.

Tennant, A. & Pallant, J. F. (2006). Unidimensionality matters! (A tale of two Smiths?) *Rasch Measurement Transactions*, 20:1, 1048-1051. Available: http://www.rasch.org/rmt/rmt201.pdf.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.