

Christine Klein-Braley

Hunting unicorns: C-Tests, cloze tests and the RIP (Removal of Information Procedure)

In 1970/71 Tuinman proposed the Removal of Information Procedure (RIP), a test in which examinees are asked to determine which words in five-word segments of texts would be most difficult for a reader to replace if they were removed. Tuinman reported that the RIP procedure – surprisingly – was uncorrelated with a variety of other language measures. This study reports a replication of Tuinman's study and produces very similar results. RIP procedures fail to produce meaningful relationships to tests aimed at utilizing the redundancy of language and texts in order to restore missing words (cloze tests, C-Tests and dictation).

1. Introduction

This study shows that empirical research does not necessarily confirm what common sense suggests must necessarily be so. I report it for two reasons. Firstly, because all too often we write up only those experiments where we found what we were looking for. Where the interrelationships we expect do not emerge, we often fail to report the data. My second reason for writing this paper is because it shows two researchers involved in hot pursuit of the unicorn. The results obtained in both studies by the two investigators involved, J. Jaap Tuinman and myself, are so fascinatingly perverse and counterintuitive that I would be delighted if they spurred on other investigators to dig deeper, and ultimately to find out just what redundancy utilization involves.

2. Tuinman's original study

In 1970/71 Tuinman proposed the use of what he described as "a reversed cloze procedure" in order to add a further dimension to the study of language processing (Tuinman 1970/71, p. 44):

"the procedure is called Removal of Information Procedure (RIP). In a RIP task, the Ss [subjects] are asked to delete those words which would be most difficult for someone to guess again. In addition, the frequency with which they are to delete words is stipulated (e.g., one word in every sentence).

The RIP is based on the assumption that language users can recognize high information words. The information value carried by a word is determined by the proportion of Ss who can fill in that word when it is deleted. An inverse relationship holds between fill-in probability and information value. Words which few Ss fill in contain much information and vice versa.

A second assumption of the RIP is that an increased familiarity with a language is reflected in an increased effectiveness in identifying high information words." (Tuinman, 1970/71, pp. 44-45).

Thus the RIP procedure asks the subjects specifically to use their knowledge of language to identify those words that would be difficult to restore in a cloze test.

In order to conduct his experiment Tuinman made cloze tests from the text which he would later use for the RIP procedure. By deleting every fifth word, starting successively for each test one word later, he constructed five parallel tests: the **basic data cloze tests**. Each test had 60 items. Thus, overall, every word in a 300-word running text was deleted in one of the tests. These five tests were administered to samples of students who were instructed to work through the cloze tests in the usual way and restore the missing words. He was therefore able to determine the empirical restoration difficulty (the *P*-values using exact scoring) for every word in the text.

For the RIP procedure itself, the same text was reprinted in its original form. However, the lines of the text were very short – each line had exactly five words – so that the text looked like a column of newsprint (see example from the Duisburg replication below). Subjects were instructed to mark in each segment the word which would be most difficult "to guess again by somebody else if he got the sentences without the words which you marked out" (Tuinman, 1970/71, p. 46).

Tuinman's scoring system for the RIP test involved assigning the *P*-value of the word the subject had marked as the score for that five-word segment of the text. If subjects correctly identified the most difficult word among the five, they received a low *P*-value; if they chose a word that was actually easier to restore according to the data obtained in the basic data cloze test, then they received a higher *P*-value as their score on that segment. Thus, subjects who were very good at identifying the words which really were hard to restore thus got a **low** total score; those who were less successful achieved **higher** scores. This would mean that if the ability to detect difficult words turned out to be related to other tests which have

higher scores if the subject performs successfully, the correlations between the RIP and these scores would be artefactually negative.

Performance on the RIP tests was correlated with scores on a classical cloze test and, for 69 of the 140 junior high school students involved, with a variety of standard test scores and IQ measures already available in school records.

Tuinman believed that the ability to identify words which would be difficult to restore if they were missing would be a specific and identifiable ability, that is, scores would differ significantly from a score derived from selecting a number between 1 and 5 at random. He also believed that the ability to perform this task would improve with age.

3. Results

Tuinman's explicit hypotheses were, in fact, confirmed by his results:

1. the mean performance of the subjects differed significantly from a mean score generated by the random deletion of words, and
2. there was a significant difference between the mean performance of subjects at different levels of linguistic development.

But implicit in his research design are other hypotheses, hypotheses which seem eminently reasonable and plausible and which one would have expected to be confirmed. For instance, he chose to administer a cloze test together with the RIP. His assumption was – presumably – that the ability to remove high information words is in some way related to the ability to make use of comprehension skills in order to complete a cloze test. In other words, he expected moderate to high correlations between the RIP scores and the cloze scores. Similarly, he probably expected to find positive relationships between the RIP scores and verbal IQ, and between vocabulary, reading, and language tests. His results are shown in Table 1.

What, in fact, emerged was entirely unexpected: "RIP performance seems uncorrelated with a variety of language related measures" (Tuinman, 1970/71, p. 50). While the RIP tests were reliable ($r_{tt(\text{Spearman-Brown})} = .73$ for the junior high school students and for the adults .63), **the correlations with all the other measures, as Table 1 shows, were to all intents and purposes zero.** Moreover, in the two factor analyses he performed, the

Table 1
Correlations between RIP tests and other measures
 (Taken from Tuinman 1970/71)

Vocabulary	.13
Reading	.15
Language	.09
Verbal IQ	.10
Non Verbal IQ	.10
IQ	.10
Cloze (<i>N</i> = 69 Junior High)	-.11
Cloze (<i>N</i> = 139 Junior High)	-.15
Cloze (<i>N</i> = 100 Adults)	-.13

RIP test was the only test loading on a unique factor. In fact, as Tuinman himself points out:

“It must be concluded that removal of information, as defined by the present RIP instrument, constitutes a performance which is both reliable and unique.” (Tuinman, 1970/71, p. 50)

Thus, the performance of these subjects on the cloze tests shows very little relationship to their understanding of text redundancy if that is what their performance on cloze tests measures.

4. The Duisburg replication

I found Tuinman's results so amazing – and so counter-intuitive – that I replicated his experiment at the University of Duisburg (Germany).

A text was selected as the basis for the RIP procedure (see full text in the Appendix). First, five basic data cloze tests deleting every fifth word were constructed. These tests were distributed at random to 128 students during a placement session in Duisburg.

The placement procedure in Duisburg always includes the DELTA test, a psychometric-structuralist test with a total of 150 multiple-choice and short-answer items aimed at measuring specific areas of grammar and vocabulary. Other tests included in the study were DICT, the dictation test normally used with DELTA, and a C-Test containing 5 superitems each with 25 deletions. Table 2 shows the basic statistics for the various tests. Note that

Table 2
Statistics for basic data cloze tests
and other tests administered in the same session

	<i>n</i>	<i>N</i>	mean	st.dev	r_{tt}	r_{tc}	r_{tc}	r_{tc}
	items	subjects			alpha	DELTA	CTEST	DICT
					or KR-20			
CLOZE1	50	24	23.68	9.33	.88	.75	.58	.73
CLOZE2	50	26	16.15	8.92	.87	.58	.53	.54
CLOZE3	50	27	16.71	8.79	.88	.89	.80	.90
CLOZE4	50	27	24.63	8.29	.87	.55	.57	.43
CLOZE5	50	24	20.82	8.06	.86	.57	.60	.54
DELTA	150	128	108.93	39.75	.94	–	.77	.94
DICT	50	128	27.10	14.59	–	.94	.71	–
C-TEST	125	128	70.71	24.09	.90	.77	–	.71

the cloze tests, although all constructed from the same text using the same deletion rate, cannot be viewed as equivalent to each other. Notice also that the C-Test outperforms all the cloze tests in terms of reliability, and all cloze tests with the exception of CLOZE3 in terms of validity.

For the next session RIP tests were constructed. Tuinman had only investigated the ability to detect words with high information values (hard to replace). I decided to investigate also the question of easily restorable words. I therefore constructed two separate RIP measures, a high information measure (HI: difficult words) and a low information measure (LO: easy words). Table 3 shows the instructions for the test and the first few lines of each part.

71 students in the placement session of WS 87/88 completed the DELTA tests, DICT, two cloze tests, a C-Test and the RIP tests.¹

Since the basic data cloze tests for analysis of RIP reliability had been administered in the previous test session, an empirical estimate for the restorability of every single word in the RIP texts was available. Two methods of calculating the RIP scores were used. The first was Tuinman's original method. A modification was made, however, in the case of high

¹ The numbers in the tables vary because some students did not process some tests.

Table 3
RIP tests used in Duisburg replication

Imagine that you want to construct a test by removing **easy** words, that is words that can easily be found and replaced in the text.

Read through the text which follows and put a ring round the word in each line that would be **easiest** to restore.

TEXT:

I am a professional man
of letters, and when I
was younger I thought a
typewriter would be convenient. I
even thought it was necessary,

...
...

Now imagine that you want to construct a test by removing difficult words, that is words that are **difficult** to replace in the text.

Read through the text which follows and put a ring round the word in each line that would be **most difficult** to restore.

TEXT:

As with telephones and typewriters,
so with cars. I obtained
my first driving licence at
the age of seventeen, having
been taught to drive in

...
...

information words. To derive the subject's score on recognizing such words the increment added was 1 minus the *P*-value. In this way, it was possible to pole the score in the normal way: a good performance in recognizing high difficulty words produces a high score. For the low information words, the *P*-values of the words were simply added together. Here, too, better performance leads to a higher score. These are the methods referred to as TUINHI or TUINLO in the tables.

A second, easier method of scoring was also used. The word which was *empirically* easiest (EMPHI), or most difficult (EMPLO) to restore in the group of five words was designated as correct. Any other word circled (to show its high or low restorability as viewed by the examinee) was counted as wrong.

Table 4 shows the basic statistics for all the tests in the study. For DELTA, CLOZE and C-TEST Cronbach's Alpha was used to calculate reliability, using each unit (subtest, text) as one superitem in the analysis. Each of the cloze tests had a possible 20 points, scored with exact scoring. Each superitem in the C-Test had a possible 25 points. The basic statistics for the four RIP procedures and their reliability coefficients are also shown in Table 4. Reliability was calculated for the Tuinman-scored procedures, TUINHI and TUINLO, using the Spearman-Brown split-half method, for the empirical procedures, EMPHI and EMPLO, using KR-20.

As the mean values of the various tests show, it is obviously easier to recognize high information words than low information words (TUINHI, EMPHI). However, the reliability coefficients show that identifying low information words (TUINLO, EMPLO) is a more reliable trait. It is interesting to note that the more complicated Tuinman scoring system produces more reliable procedures: the variation in the scores for the individual words in the basic data cloze tests is reflected in the Tuinman scoring system but not in the empirical system. The easier method of scoring obviously lowers reliability. Even so, the results confirm Tuinman's conclusions: for research purposes the RIP procedures are acceptably reliable measures of whatever they measure.

Tables 5 and 6 show the correlations between the two differently scored RIP measures and the other tests run in the same session.

Naturally my assumptions were the same as those I have attributed to Tuinman. The ability to recognize high and low information words must – surely – be related to the ability to make use of the redundancy of natural

Table 4
Basic statistics of the tests in the study (N between 66 and 71)

	n items	mean	st.dev	r_{tt} (Spearman-Brown)	r_{tt} (alpha)
DICT	50	30.81	13.39	—	—
CLOZE	40	10.73	4.73	—	.66
DELTA	150	83.16	22.57	—	.90
C-TEST	100	51.14	16.59	—	.86
TUINHI	25	15.39	4.56	.80	—
TUINLO	25	10.27	5.09	.86	—
EMPHI	25	13.68	3.76	.58	.63
EMPLO	25	8.49	3.93	.68	.78

Table 5
Correlations between RIP tests using Tuinman's scoring and other tests (N between 66 and 71)

Test	TUINLO	TUINHI	DICT	CLOZE1	CLOZE2	DELTA
TUINHI	.17					
DICT	.08	-.04				
CLOZE1	.06	-.01	.59**			
CLOZE2	.11	.03	.43**	.58**		
DELTA	.07	.00	.71**	.71**	.62**	
C-TEST	.06	-.05	.57**	.71**	.62**	.70**

** : two-tailed significance < .001

Table 6
Correlations between RIP tests using empirical scoring and other tests (N between 66 and 71)

Test	EMPHI	EMPLO	DICT	CLOZE1	CLOZE2	DELTA
EMPLO	.08					
DICT	.05	.03				
CLOZE1	.13	.16	.59**			
CLOZE2	.16	.25	.43**	.58**		
DELTA	.11	.15	.71**	.71**	.62**	
CTEST	.12	.18	.57**	.71**	.62**	.70**

** : two-tailed significance < .001

language as expressed in successful processing of reduced redundancy tests. This ought to be reflected in moderate to high correlations between the RIP measures, DICT, cloze tests and C-Tests.

The cloze and C-Tests do indeed correlate with each other, and with DELTA and DICT at values between .43 and .71. The highest correlations in the table are those between DELTA and DICT, CLOZE1 and DELTA, CLOZE1 and C-TEST at .71. Interestingly, each of the cloze tests has a higher correlation with the C-Test (CLOZE1 .71; CLOZE2 .62) than with the other cloze test (.58). **But none of the three redundancy utilization measures has a higher correlation with any of the RIP procedures than .25.** Even more interestingly, the correlation between TUINHI and TUINLO is only .17, and that between EMPHI and EMPLO .08.

These results confirm those reported by Tuinman in almost all respects (no intelligence test was included in Duisburg). He is quite right: RIP procedure does constitute a performance which is both reliable and unique. Finding high or low information words is not related to tests aimed at utilizing redundancy for word restoration. Moreover, **there is also no relationship between the ability to find high information words, and that to find low information words either!**

The precise relationships between the two different types of scoring and the two different operations of detecting high and low information words are shown in Table 7. Here, at least, expected relationships emerge: the two high information measures (EMPHI and TUINHI) produce a correlation coefficient of .61 while the two low information measures (EMPLO and TUINLO) produce a coefficient of .85.

Table 7
Correlations between empirical and Tuinman's scoring for RIP tests (N = 66)

	EMPHI	EMPLO	TUINHI
EMPLO	.08		
TUINHI	.61**	.07	
TUINLO	-.03	.85**	.02

** : two-tailed significance < .001

More extensive statistical analysis reveals little more information. Factor analyses performed over the set of tests show clearly that all the overt language tests² load on Factor 1, which is virtually the only interpretable factor in the matrix. The only language test to perform consistently at a lower level than the others is CLOZE2, which has a lower initial and final communality in all solutions and is the lowest language test loading on the general factor. The RIP tests, whichever variant is used, have low initial communality and low loadings on Factor 1.

The RIP tests are the only tests loading on Factor 2, which is thus a RIP factor, as Table 8 shows.

Table 8
Initial solution for factor analysis using Tuinman's scoring

	FACTOR 1	FACTOR 2
TUINHI	-.01	.46
TUINLO	.10	.35
DICT	.77	-.06
CLOZE1	.80	-.02
CLOZE2	.68	.11
DELTA	.88	-.02
CTEST	.81	-.04

For the construct validation of the C-Test these results produce no new insights.

5. Conclusions

Despite Tuinman's report of his results I was convinced at the beginning of this study that there must be a relationship between the RIP procedures and the redundancy utilization language tests. There is no relationship.

The tests are acceptably reliable if the easier EMP procedure is used for scoring, assigning one point for identifying the easiest or most difficult word in the group of five as determined by the data of the basic cloze tests

² I use the term **overt** because identifying words of high or low information must also be a language test of some kind.

administered to a different group in a previous session. If Tuinman's own scoring procedure is used, assigning points on the basis of the *P*-values of the words concerned, this is more complicated but produces even more reliable results. This applies to both types of identification process, whether the examinee is expected to find high information or low information words. However, these two processes also have nothing in common: for all practical purposes the correlation between them is zero.

So, we have a reliable procedure, but what does it measure? If it is a valid test, what is it a valid test of? Douglas K. Stevenson was once heard to mutter at a language testing meeting that you only need to leave language tests together in the same room for them to start correlating with each other. Here is a test, apparently related in some way to language processing, which refuses to correlate with anything. Truly, the RIP procedure is a magnificent enigma, and I offer it herewith to other researchers for their delectation.

References

Tuinman, J. Jaap. (1970/71). The Removal of Information Procedure (RIP). A first analysis. *Journal of Reading Behavior*, 3(2), 44-50.

Appendix: The text

I am a professional man of letters, and when I was younger I thought a typewriter would be convenient. I even thought it was necessary, and that editors and publishers would expect anything sent to them to be typewritten. So I bought a typewriter and taught myself to type. But I did not enjoy typing. I enjoy forming letters or words with a pen, and I never could enjoy tapping the keys of a typewriter. The fact is, I am not mechanically-minded and the typewriter is a machine. I have never really been driven to machines. And machines do not like me. When I touch them they tend to break down, get jammed, catch fire or blow up.

As with telephones and typewriters, so with cars. I obtained my first driving licence at the age of seventeen, having been taught to drive in the rush hours of the busy city of Johannesburg. I needed the car for use in another part of Africa where in those days there was hardly any motor traffic. The actual process of driving soon became automatic, and my sole idea was to get from one place to another as soon as possible. I therefore drove fast, and within a week or two the speedometer was broken. I never had it mended. I was not a reckless driver. I did not lose control of the car, even on rocky or sandy tracks or driving with chains through deep mud. I never killed or injured anybody. But I was bored.