Moosbrugger, Helfried & Müller, Hans. (1981). Ein klassisches latent-additives Testmodell (KLA). In Wolfgang Michaelis (Hrsg.), *Bericht über den 32. Kongreß der Deutschen Gesellschaft für Psychologie in Zürich 1980* (Bd. 2, S. 482-486). Göttingen: Hogrefe.

Raatz, Ulrich & Klein-Braley, Christine. (1992). *CT-D4. Schulleistungstest Deutsch für 4. Klassen. Test und Beiheft mit Anleitung und Normentabellen.* Weinheim: Beltz.

Roos, Undine. (1994). The C-Test in Japanese. In Grotjahn (1994), 61-113.

Roos, Undine. (1995). *Ein C-Test für Lerner der japanischen Sprache: Entwicklung, Erprobung und Validierung.* Bochum: AKS.

Sigott, Günther. (1995). The C-Test: Some factors of difficulty. *Arbeiten aus Anglistik und Amerikanistik, 20*, 43-53.

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

Stemmer, Brigitte. (1992). An alternative approach to C-test validation. In Grotjahn (1992), 97-144.

Süßmilch, Edgar. (1984). Language testing with immigrant children. In Terry Culhane, Christine Klein-Braley & Douglas K. Stevenson (Hrsg.), *Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the IUS, Colchester, October 1983* (S. 167-176). Colchester: University of Essex, Department of Language and Linguistics.

Tuinman, J. Jaap. (1970-71). The removal of information procedure (RIP). *Journal of Reading Behavior, 3*, 44-50.

Grotjahn, Rüdiger. (Hrsg.). (1996). Der C-Test. Theoretische Grundlagen und praktische Anwendungen (Bd. 3, S. 23–94). Bochum: Brockmeyer

Christine Klein-Braley

Towards a theory of C-Test processing

This paper makes a contribution to the vexed question of the psycholinguistic validation of the C-Test. C-Tests have been shown to have empirical validity. However it is still unclear what psycholinguistic mechanisms underlie test performance. This article investigates responses given by over 200 individuals (German university freshers with an average age of 19 years; English schoolchildren with an average age of 12) to individual blanks in 30 different C-Test texts in English. In the study, test statistics including difficulty indices, item discrimination indices, inter-item correlations, reliability and validity coefficients, and factor analysis are used as tools to investigate the linguistic response behaviour of the test subjects. The analysis deals with correct and incorrect responses, compares the performance of the English and the German groups, seeks to detect specific strategies and approaches, looks at solution strategies for items on the word, sentence and text level. It is shown that C-Test performance can be meaningfully related to the concept of general language proficiency.

1. C-Tests as tests

In our research into C-Test performance Raatz and I have been able to show that tests based on the C-principle function as proficiency tests for a variety of different groups, in particular for children learning their own language (L1 learners), for children and adults learning a second language in the country in which it is spoken (L2 learners), and for foreign language learners (LF learners) (cf. Klein-Braley & Raatz, 1985; Raatz & Klein-Braley, 1992). Our own original experiments referred only to the languages English and German, but very soon Süßmilch (1984, 1985) was able to demonstrate that C-Tests also worked in Greek, Serbo-Croat and Turkish; Grotjahn and the Bochum group (Feldmann, Grotjahn & Stemmer, 1986; Grotjahn, 1986; Grotjahn & Stemmer, 1985) obtained satisfactory results in French and Spanish (cf. also Lütticken, 1985); and Cohen, Segal & Weiss Bar-Siman-Tov (1985) were also able to develop highly reliable tests in Hebrew for new immigrants to Israel. In the past ten years C-Tests have become standard instruments in a variety of countries and for a variety of purposes. The papers in this and its companion volume (Grotjahn, 1992) demonstrate the success of this test construction technique.

In virtually all the studies thus far reported the C-Tests have been shown to be highly reliable, with alpha coefficients very often higher than .9, and to have high correlations with whatever other measure was used to represent language proficiency: teacher ratings or judgments, self-assessment procedures, and other language tests and language testing procedures. These validity coefficients have regularly reached .7 and higher. Such high validity coefficients are unusual for any type of test. In addition, investigations have been conducted into other aspects of C-Test performance: retest reliability has been shown to be high even over long intervals (Süßmilch, 1985), the factorial structure of German C-Tests has been explored (Raatz, 1984, 1985d) and shown to conform to the previously hypothesized patterns. The theoretically expected differences in the performances of different ability groups have been confirmed to be in the expected directions (Raatz, 1985b). Hopkins (1985) and Rumpel (1985) have contributed theoretical underpinnings to the rationale behind the tests. For English and German it has been possible to set up regression equations using indices derived from text statistics to predict empirical difficulty for specific groups (Klein-Braley, 1985b; Klein-Braley, 1994). Thus, in terms of classical test theory, C-Tests are very satisfactory testing instruments: their objectivity, reliability and **empirical** validity are not in doubt, and considerable advances have been made towards demonstrating essential aspects of **construct** validity (cf. Klein-Braley, 1985a, 1985c; Raatz, 1985a, 1985d).

2. What do C-Tests measure?

One question remains to be answered: what exactly do C-Tests measure in terms of language processing? In one sense, of course, the answer to this question is irrelevant. Let me illustrate this by an example. Supposing it was possible to demonstrate high correlations (around .7 or higher) between the distance a stone can be lobbed and the probability that the stone-lobber would cause a road accident in the next calender year. In view of the magnitude of the correlation coefficient there would be every justification for society to take very strong measures to ensure that those who are high risks (the far lobbers!) were prevented in some suitable way from causing the potential accidents. The reason for the high correlation need not be known; it might be posited in some intervening variable, for instance the personality structure of the individual involved. But whatever the cause of the relationship, if it is stable, then decisions can be based on it. To

move to the real world: the statistical relationship between the likelihood of dying of lung cancer and cigarette smoking is lower than .7. Nevertheless each packet of cigarettes carries a government health warning, and the European Community is currently considering a ban on all forms of cigarette advertising.

So, if the C-Test works, we can use it, whatever it actually measures.

On the other hand, given that the tests are so successful, the applied linguist and language tester would like to find out what is actually going on **psycholinguistically** when subjects complete a C-Test. Can we open the black box and watch language processing actually taking place? There is no doubt that a satisfactory account of the mental processes underlying test performance would be a very useful contribution to construct validation.

3. Investigation of test-taking strategies

For the investigation of test-taking processes Grotjahn (1986b; see also Feldmann et al., 1986) suggests three possible approaches: statistical item analysis, text linguistic item analysis and analysis of individual performance. The research in Duisburg conducted by Raatz and myself has primarily used the classical techniques of test and item analysis. The Bochum group under the leadership of Grotjahn have concentrated their efforts on the third approach listed above: investigations of individual performance while subjects are completing C-Tests. This study will present evidence derived from text and test linguistic and statistical item analysis.

3.1 Investigation of test-taking strategies: think-aloud protocols

The researchers in Bochum have primarily used think-aloud protocols in order to look inside the black box (see e.g. Feldmann et al., 1986; Feldmann & Stemmer, 1987; Grotjahn, 1986b, 1987b; Stemmer, 1991, 1992). There is no doubt that analysis of these protocols has substantially increased our understanding of what goes on inside the subject during test-taking. The methodology involves asking subjects to verbalise their thoughts while taking the tests. After the test has been completed, the tapes are then replayed to the subject who can answer any questions put by the investigator, clarify remaining uncertainties, or add comments if necessary. This retrospective or communicative validation of the think-aloud protocols is important, although, as Feldmann et al. (1986, p. 340) point out themselves, the researcher is still dependent on the cognitions of the individual subject, which

may be only subjectively true, while objectively false. Only around 50% of the blank-completion processes appear to be open to retrospective analysis, or talking-aloud techniques. The other solutions – whether correct or incorrect – simply "appear", apparently as the result of automatic processing.

There are problems in using this technique. One major drawback for anyone trained in a quantitative tradition is the small number of subjects involved. Feldmann et al. (1986) report results for a total of 20 subjects in all (ten each in Spanish and French). Stemmer's (1991) study reports data for 30 subjects.

This is not surprising in view of the immense labour involved in collecting the data: each C-Test session lasts three to four hours including the follow-up, and the data then have to be transcribed. Thus, up to 20 researcher hours can be invested in the results for just one subject. However these small numbers mean that generalisations are being made on a very slender data base, and the model does not provide for individual differences in processing strategies. But it would only need one subject to behave in an idiosyncratic manner to lead to very odd conclusions being drawn from the protocols.

Another problem is that the results may be affected by a very considerable amount of reactivity. The think-aloud protocols concentrate on cognitive processes, but an interaction between the process of test-taking and the process of commenting on test-taking cannot be excluded.

What seems to me to be most problematic, however, is the question of task authenticity. In terms of modelling C-Test performance, the think-aloud experiment is very far from real life. In Duisburg, for instance, examinees are given 25 minutes to complete 6 C-Test texts with 25 blanks each. They work in a large room in the company of other examinees in the context of a placement testing session. Their performance is pragmatically meaningful since their scores will be used, together with those on the other subtests, to determine their level of placement.

Stemmer's subjects, on the other hand, were tested individually. She stresses that she made great efforts to give them the feeling that they were "competent collaborators" (Stemmer, 1991, p. 55). The introduction, during which they were played a tape of an extract from a think-aloud session, lasted about 20 minutes. They then processed three French texts in a C-Test format, thinking aloud, in a room with a tape recorder running. This took around 30 minutes. Then they went through a debriefing session, listening to the tape with the investigator, making comments and answering questions.

How far the results derived from such sessions can answer the question of what happens when C-Tests are processed under "normal" conditions is still very much open to question. Unfortunately, as Stemmer herself points out, in her study there was no "normal" control group. This means that we have no way of knowing, for instance, whether the tests done under think-aloud conditions were similar in difficulty to "normal" tests, whether the solutions offered were similar to those offered by "normal" subjects and so on. One of Stemmer's main conclusions is that "the strategic behaviour inferred from the verbal protocols can be interpreted as reflecting predominantly lower level processing in C-Test solving" (Stemmer, 1991, p. 301). This may be true, though I doubt it. However one of the major differences between speech and writing is that in speaking we are involved in real-time on-line language processing in a constant linear flow, whereas in reading and writing we are concerned with conserved fixed language units enabling recursive processing to take place. It may well be that the request to "think aloud" encourages the examinee to procede in a linear fashion which is not typical of normal C-Test processing. Quite clearly, then, think-aloud techniques, while informative, are not conclusive evidence of what goes on in C-Test processing.

3.2 Investigation of test-taking strategies: non-reactive approaches

Is there any way of getting inside the black box and yet avoiding the problem of reactivity? It seems to me that the other two approaches suggested by Grotjahn (1986b) offer precisely this possibility. They involve an ex post facto analysis of the statistical indices of individual items and their interrelationships, with the aim of interpreting these in terms of linguistic phenomena. Secondly it is possible to perform an inspection of the test-taking behaviour of the test subjects as evidenced by the test scripts. Again, the aim is to detect the psycholinguistic processes involved in test taking. I have already published a preliminary investigation based on procedures of this type (cf. Klein-Braley, 1985a); in the present investigation, however, the data base is much larger and the approach has been more systematic.[1]

[1] See Germann & Grotjahn (1996) for a related approach using a computer to

4. This study: basic data

I base my conclusions on data drawn from over 200 subjects who completed in all a total of 30 different English C-Test texts. Tables 1 and 2 summarize the basic data. Around one third of the group are English 11- to 13-year-olds from Wales Comprehensive School in South Yorkshire.[2] For these subjects no validation data are available. For the German subjects, who completed some of the same texts in the context of placement sessions at the University of Duisburg, extensive demographic and other data are available, including the results of the various subtests of the DELTA test in the winter semester (WS) 90/91, a test whose own validity has been extensively investigated and been shown to be satisfactory.

4.1 Statistical analysis

One major source of information is provided by the statistical analysis of test responses, blank for blank.

So far as statistical item analysis for purposes of **test quality control** is concerned, Raatz and I have stipulated item analysis on the super-item level, using each text in the C-Test as one item. We have repeatedly warned against performing item analysis on the basis of individual deletions in C-Tests since it seemed obvious to us that the individual blanks – successive deletions in the same text – in a C-Test must be linguistically dependent on each other. Traditional methods of item analysis make the assumption that items are statistically independent. The justification for this warning has been amply confirmed by the research reported here. Individual inter-item correlations between C-Test deletions are frequently very high, and not infrequently reach 1.00 between successive items. In addition, mean item intercorrelations regularly reach levels between .4 and .5.

However there is no reason why item analysis should not be performed over individual deletions as a **research technique**, and to complete the triangulation recommended by Grotjahn (1986b, 1987b), this paper will report the results of statistical item analysis performed on the C-Tests.

---

record the test-taking behaviour of their subjects.

[2] My thanks to the teachers and pupils who cooperated in the experiment, in particular to Ms Judith Cole.

Table 1
Basic data German group

Test	FORM	N	MEAN	St.Dev	P	r_{tt}	r_{DEL}	r_{dict}	r_{tot}
CARS		120	17.91	4.86	71	.85	.69	.73	.72
LITERATURE		118	13.12	3.93	52	.78	.60	.59	.61
CARS	1	23	16.61	4.92	63	.85	.77	.81	.79
	2	21	17.23	5.29	67	.87	.53	.64	.57
	3	22	16.90	4.87	64	.84	.79	.82	.82
	4	25	18.60	4.62	70	.85	.72	.70	.72
	5	29	19.62	4.38	75	.85	.62	.62	.63
LITERATURE	1	22	11.81	3.43	53	.70	.56	.59	.57
	2	21	12.71	3.84	55	.77	.93	.86	.93
	3	21	11.33	4.00	49	.76	.44	.42	.44
	4	25	11.88	3.96	51	.78	.69	.66	.68
	5	25	13.48	4.05	58	.79	.54	.54	.54
HEART	1	23	17.39	3.92	70	.77	.86	.85	.86
COURTS	1	23	13.39	3.58	54	.65	.62	.63	.62
COMPUTER	1	23	15.08	5.97	60	.90	.76	.76	.76
ALL C-TESTS (5 items)			74.28		60	.86*	.81	.81	.82
JUPITER	2	21	15.90	5.42	64	.71	.66	.63	.67
BLOOD	2	21	17.09	4.44	68	.78	.76	.69	.76
BACON	2	21	13.38	5.45	54	.78	.82	.76	.82
ALL C-TESTS (5 items)			76.31		61	.90*	.86	.86	.88
ASTROLOGY	3	22	15.40	3.83	62	.73	.74	.83	.79
SOCIOLING	3	22	16.40	5.17	66	.86	.74	.76	.77
HORMONES	3	22	16.04	4.22	64	.76	.85	.77	.84
ALL C-TESTS (5 items)			76.07		61	.91*	.83	.84	.85
PROBLEMS	4	25	17.48	4.04	70	.80	.77	.73	.76
CANALS	4	25	15.80	5.13	60	.87	.66	.62	.65
FRENCH	4	25	13.16	4.49	53	.80	.55	.46	.53
ALL C-TESTS (5 items)			76.92		62	.88*	.81	.76	.80
WORK	5	29	19.82	4.81	79	.89	.72	.72	.71
CUISINE	5	29	18.66	4.03	75	.83	.70	.73	.91
FREESPEECH	5	29	15.58	4.66	63	.84	.61	.61	.62
ALL C-TESTS (5 items)			87.16		69	.88*	.78	.75	.79
DICT		118	28.65	13.98	57				
GRA1		119	16.73	3.93	67				
VOC3		119	27.97	7.79	56				
GRA2		120	13.64	5.18	55				
GRA3		120	15.20	4.55	61				
VOC1		120	14.35	5.27	57				
JOINTC		118	31.09	7.79	62				

* calculated by Cronbach's alpha; all other reliabilities = KR-20

Table 2
Basic data English group

Test	FORM	*N*	MEAN	St.Dev	*P*	r_{tt}
CARS	A	24	20.00	2.52	80	.52
CUISINE		24	19.08	2.13	76	.61
RICHARD		24	18.74	2.56	75	.59
CHIMPS		24	18.75	2.67	75	.55
ALL C-TESTS (4 items)			76.52	7.34	77	.75*
PETS	B	24	21.25	2.77	85	.72
AMERICAN		24	18.37	4.59	73	.73
CANAL		24	17.54	6.42	70	.93
CORNWALL		24	18.29	6.49	73	.93
ALL C-TESTS (4 items)			75.45	17.16	75	.89*
TELEPHONE	C	24	19.38	5.56	78	.84
DIET		24	16.75	4.96	67	.89
WINDMILLS		24	16.08	5.04	64	.85
CONCRETE		14	13.20		53	.92
FIRST 3 CTESTS			52.21	11.65		.67*
ALL C-TESTS (4 items)			65.41	17.54	65	.83*
BEHAVIOUR	D	27	17.55	5.82	70	.89
RIVERS		27	17.81	5.10	71	.86
FLIGHT		27	15.25	6.29	61	.91
DECORATING		27	18.59	5.97	74	.91
ALL C-TESTS (4 items)			69.20	21.04	69	.93*

* calculated by Cronbach's alpha; all other reliabilities = KR-20

4.2 Text script analysis

A further rich source of information is provided in the responses to C-Tests as revealed by the test papers themselves (cf. Klein-Braley, 1985a). The language processing techniques of individual test subjects and groups of subjects as shown by their responses to C-Test items demonstrate quite clearly why C-Tests can legitimately claim to represent the concept of general language proficiency. I shall show how test-taking behaviour – as evidenced by the entries on the test scripts – enables us to make reasonable guesses about the nature of the language processing underlying the – correct or incorrect – responses.

4.3 Possible over-interpretation of data?

The greatest danger inherent in an ex-post-facto analysis of test scripts is the possible over-interpretation of what one finds there. For instance, I contend that crossings out reveal reprocessing by the test subject and that incorrect responses can be interpreted as revealing specific psycholinguistic processes. I also contend that incorrect response behaviour is in many cases less trivial and careless than it seems.

To provide just one example here, given a context such as *Wine is generally cheaper than in a British hotel*, I shall argue that those subjects who originally wrote ho**me**, crossed it out and replaced it with ho**tel**, engaged in reprocessing, probably as a result of their pragmatic knowledge that wine is not normally sold in homes, not even in British ones, despite the price of alcohol in England. I think this is a legitimate conclusion; the reader may disagree. Furthermore I would claim that the reprocessing demonstrated here takes place at a very high level of text processing, the pragmatic level, so that although the change affects only one word, the difference in overall performance between those who left the word *home* standing and those who changed it to *hotel* is a meaningful one which can be interpreted in terms of the individual's standing on the language proficiency continuum.

4.4 Future research: post-hoc validation of test responses?

There is no reason why the two approaches used in Bochum and Duisburg could not be combined (cf. Little & Singleton, 1992). A cycle of post-validation by consultation of the test subjects as used in the think-aloud research could be performed after normal test taking. The experimental set-up would involve collecting data for a set of texts and developing hypotheses about language processing from them, then administering the same texts to a second group of examinees. These subjects would be questioned about their solutions. Finally, the original hypotheses, the statistical results and the respondent's own interpretation could be matched. This would, of course, reintroduce an element of reactivity, but the reactivity would be ex post facto and not on-line, which makes it less problematical.

5. Analytical procedures used in this study

For all the tests involved in this investigation normal statistical analysis on the super-item basis was performed; in addition, for every text an individual

item analysis was calculated on the basis of the deletions in the texts. For some texts the responses of all respondents were transcribed, for others notes were taken of particularly interesting phenomena. The selection of the depth of detail for the analysis of individual texts was determined by a variety of considerations. For instance, some texts were completed by more than one group, which made it possible to compare responses, others had been completed by very large groups of subjects, and so on. Tables 1 and 2 show the basic empirical data.

It is not possible without exceeding the limits available to me in this publication to reprint all the data available; however, anyone interested in this approach will easily be able to obtain similar data, and will, I predict, quickly detect that the type of mechanisms I describe here operate in any set of C-Tests. For the reader's orientation Tables 3 and 4 present one full set of C-Test responses: the English group's responses to the GERMAN CUISINE text. Other data sets are similar.

> *German cuisine seems very much to the English taste. Emphasis i(1)***s** *on me(2)***at** *– less s(3)***o** *on fi(4)***sh** *and veget(5)***ables**. *A salad of(6)***ten** *accompanies a me(7)***al**. *Wine i(8)***s** *generally che(9)***per** *than i(10)***n** *a British ho(11)***tel**. *If y(12)***ou** *are eat(13)***ing** *out t(14)***he** *choice o(15)***f** *food i(16)***s** *varied i(17)***n** *both Ger(18)***many** *and Aus(19)***tria**, *ranging fr(20)***om** *local specia(21)***lities** *to international cui(22)***sine**. *Prices a(23)***re** *generally lo(24)***wer** *than i(25)***n** *Britain.*

Table 4 shows the same response behaviour ordered according to test scores.

Table 3
Responses of English group, text GERMAN CUISINE

Emphasis	i_(1)_	on me_(2)_,	less s_(3)_	on fi_(4)_	and veget_(5)_
1.	–	meals	–	fibre	vegetables
2.	it	meat	so	fish	vegetables
3.	is	meals	on	fitness	vegetables
4.	is	meat	seasoning	fish	vegetables
5.	is	meat	sugar	fish	vegetables
6.	is	meals	sugar	fish	vegetables
7.	is	meat	so	fish	vegetables
8.	is	meat	seasoning	fish	vegetables
9.	it	meat	seasoning	fish	vegetables
10.	is	meat	salt	fish	vegetables
11.	is	meat	salad	fish	vegetarian
12.	–	meals	salad	fish	vegetables
13.	–	meal	–	–	vegetables
14.	is	meat	sauce	fish	vegetables
15.	is	meat	so	fish	vegetables
16.	is	meat	salt	fish	vegetables
17.	is	meals	sauces	filings	vegetables
18.	is	meals	sauce	filings	vegetables
19.	is	meal	salt	fish	vegetables
20.	is	meat	salad	fish	vegetables
21.	is	medium	sone	fighting	vegetive
22.	is	meals	salt	fish	vegetables
23.	is	meat	salt	fish	vegetables
24.	is	meat	so	fish	vegetables

A salad	of_(6)_	accompanies a	me_(7)_.	Wine	i_(8)_	generally
1.	often		meal		is	
2.	often		meal		is	
3.	often		meal		is	
4.	often		meal		is	
5.	often		meal		is	
6.	often		meal		is	
7.	often		meal		is	
8.	–		meat		is	
9.	often		meal		is	
10.	often		meat		is	
11.	often		meal		is	
12.	often		meat		is	

13.	of tomatoes	meal	is
14.	often	meal	is
15.	often	meat	is
16.	often	meat	is
17.	often	meal	is
18.	often	meal	is
19.	–	meal	is
20.	often	meal	is
21.	often	meal	is
22.	often	meal	is
23.	often	meal	is
24.	of vegetables	mealfull	is

che_(9)_ than i_(10) a British ho_(11). If y_(12) are eat_(13)

1.	cheap	in	holiday	you	eating
2.	cheaper	in	hotel	you	eating
3.	cheap	is	hotel	you	eating
4.	cheaper	in	hotel	you	eating
5.	cheaper	in	hotel	you	eating
6.	cheaper	in	hotel	you	eating
7.	cheaper	in	house	you	eating
8.	cheaper	in	home	you	eating
9.	cheaper	in	home	you	–
10.	cheaper	in	hotel	you	eating
11.	cheaper	in	home	you	eating
12.	checked	in	home	you	eating
13.	cheaper	indias	–	you	eating
14.	cheaper	in	hotel	you	eating
15.	cheaper	in	home	you	eating
16.	cheaper	in	home	you	eating
17.	cheaper	in	holiday	you	eating
18.	cheque	is	holiday	you	eating
19.	cheaper	in	hotel	you	eating
20.	cheaper	in	home	you	eating
21.	cheaper	in	house	you	eating
22.	cheaper	in	holiday	you	eating
23.	cheaper	in	home	you	eating
24.	cheaper	in	hotel	you	eating

out, t_(14) choice o_(15) food i_(16) varied i_(17) both

1.	the	of	is	in
2.	the	of	is	in
3.	the	of	is	in
4.	the	of	is	it is
5.	the	of	is	in
6.	the	of	is	in
7.	the	of	is	is
8.	the	of	is	in
9.	the	of	is	in
10.	then	–	is	in
11.	top	of	is	in
12.	the	of	is	in
13.	the	of	is	in
14.	the	of	is	in
15.	the	of	is	in
16.	the	of	is	in
17.	the	of	is	in
18.	the	of	is	in
19.	the	of	is	in
20.	the	of	is	in
21.	the	of	is	in
22.	the	of	is	in
23.	try	of	in	in
24.	the	of	is	in

Ger_(18) and Aus_(19), ranging fr_(20) local specia_(21)

1.	Germany	Austria	from	specialities
2.	German	Austrian	from	specials
3.	Germany	Austria	from	special
4.	Germany	Australian	from	specialities
5.	Germany	Austria	from	specialities
6.	Germany	Austria	from	specialities
7.	German	Austrian	fruit	specials
8.	Gernerations	Australia	from	special
9.	German	Austrian	from	special
10.	German	Austrail	from	speciality
11.	German	Australian	from	special
12.	German	Australian	from	specialists
13.	Germany	Austria	from	specials
14.	German	Australian	from	specials

15.	German	Austrian	from	specials
16.	German	Australian	from	specialise
17.	German	Australia	fruit	specials
18.	German	Australia	fruit	specials
19.	Germany	Austria	fried	special
20.	Germany	Australuan	from	specialities
21.	Germany	Austria	from	special
22.	German	Australian	from	specialities
23.	German	Australian	from	specialities
24.	German	Austural	from	speciality

to international cui_(22). Prices a_(23) generally lo_(24)

1.	cuisine	are	lower
2.	cuisine	are	lower
3.	cuisine	are	low
4.	cuisine	are	lower
5.	cuisine	are	lower
6.	cuisine	are	lower
7.	cuisine	are	low
8.	cuisine	are	low
9.	cuisine	nd	lower
10.	cuisine	are	lower
11.	cuisine	are	lower
12.	cuisine	are	lower
13.	cuisine	–	lower
14.	cuisine	are	lower
15.	cuisine	are	lower
16.	cuisine	are	lower
17.	cuisine	are	lower
18.	cuisine	are	lower
19.	cuisine	are	lower
20.	cuisine	are	lower
21.	cuisine	are	low
22.	cuisine's	are	lower
23.	cuisine	are	lower
24.	cuit	are	lower

(all 24 subjects responded correctly with than in (25) Britain)

Table 4
Results of English group, C-Test CUISINE
arranged by score

Total	Item																									Subj.
										10											20						
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	#
16				1	1	1		1		1		1	1	1	1	1	1			1		1	1	1	1	12
16					1		1	1	1			1	1	1	1	1	1	1	1	1		1		1	1	13
17	1	1		1	1			1	1	1		1	1	1	1	1	1			1		1	1		1	8
17		1		1	1	1	1	1	1	1		1		1	1	1	1			1		1		1	1	9
17	1				1	1	1	1	1	1		1	1	1	1	1	1		1			1	1		1	17
17	1				1	1	1	1	1			1	1	1	1	1	1		1			1	1	1	1	18
18	1				1	1	1	1			1	1	1	1	1	1	1	1	1	1		1	1		1	3
18	1	1		1	1	1		1	1	1	1	1	1			1	1			1		1	1	1	1	10
18	1	1		1		1	1	1	1	1		1	1		1	1	1			1		1	1	1	1	11
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						1	1		1	7
19	1	1		1	1	1		1	1	1		1	1	1	1	1	1			1		1	1	1	1	16
19	1			1	1		1	1		1	1	1	1	1	1	1	1	1	1			1	1	1	1	19
19	1					1	1	1	1	1		1	1	1	1	1	1	1	1	1		1	1	1	1	21
19	1	1		1	1	1	1	1	1	1		1	1		1		1			1	1	1	1	1	1	22
19	1	1		1	1	1	1	1	1	1		1	1		1		1			1	1	1	1	1	1	23
19	1	1	1	1	1			1	1	1	1	1	1	1	1	1	1			1			1	1	1	24
20	1				1	1	1	1		1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
21		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			1		1	1	1	1	2
21	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1			1		1	1	1	1	14
21	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1			1		1	1	1	1	15
22	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1		1		1	1	1	1	1	1	4
23	1			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6
23	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	20
24	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	5

6. Response behaviour

There are three possible basic types of response behaviour to a C-Test blank. The examinee can give **no** answer at all, a **wrong** answer or a **correct** answer.

6.1 Unsolved blanks: early closure and narrow focus

There seems no way of interpreting unsolved blanks except as a breakdown in text processing. Statistically we can compare examinee behaviour over

the same blank and discover how difficult individual blanks are for the group overall, but a blank on its own cannot normally reveal why processing broke down at this particular point in the text. What *is* frequently associated with blanks is preceding **early closure**: the subject completes a sentence or part of a sentence in a way which seems syntactically and semantically correct at that point. However, because the unit has now been apparently completed, the subject cannot make any sense at all of the next blanks. For instance, in FREE SPEECH we find the following examples of early closure:

FREE SPEECH

*It has been suggested that the survival of free speech in England is partly the result of stupidity. This m(1)***ay** *be tr(2)***ue**. *But t(3)***he** *instincts a(4)***nd** *traditions o(5)***n** *which t(6)***he** *English re(7)***ly** *served th(8)***em** *best wh(9)***en** *they we(10)***re** *an except(11)***ionally** *fortunate peo(12)***ple**, *protected b(13)***y** *geography fr(14)***om** *major disa(15)***sters** ...

1. Subject 8
   But the instincts **are** *traditions* <early closure> *o_____ which t_____ English really served them best*[3];
2. Subject 12
   But **true** *instincts* **are** *traditions* <early closure> *o_____ which t_____ English*;
3. Subject 19
   ... *the English really served* **the** *best* **where** *they* **were** *an* **exception** <early closure> *fortunate peo_____*;

and in PROBLEMS:

*How do people solve problems? The sh(1)***ort** *and sim(2)***ple** *answer i(3)***s** *that psycho(4)***logists** *do n(5)***ot** *know. Th(6)***ere** *is n(7)***o** *accepted the(8)***ory** *of prob(9)***lem** *solving. Pa(10)***rt** *of t(11)***he** *difficulty i(12)***s** *that prob(13)***lems** *come i(14)***n** *all sha(15)***pes** *and si(16)***zes** ...

4. Subject 4
   Part of the difficulty is that problems come in all shapes and **since** *from* <early closure> _____ *riddles* ...;

[3] The second half of this solution shows textual reinterpretation.

5. Subject 7
   Part of the difficulty is that probably <early closure> *come i_____ sha_____ and si_____*;
6. Subject 11
   This is now accepted that <early closure> *of pro_____ solving*.

Not all texts provoke or permit early closure.

Much more frequent is **narrow focus**, which, naturally, is also associated with blanks: often individual units are supplied for an item surrounded with blanks. Such items fit into the local context, but not into the macro-context. Many examples could be given. For instance, in this extract from CAR as processed by Subject 15 **trading condition** is in itself a plausible unit, but not in the context given:

what y_____ are to_____ about **trading** *condition a_____ value o_____ the c_____*.

6.2 Different but correct responses

In some, generally in few, cases alternative responses can be viewed as correct. In HORMONES there is no alternative response I would accept. One might quarrel about *sexual* (# 5) for *sex hormones*, but in my view this is a fixed term and not open to variation. However in SOCIOLINGUISTICS I would accept *geographic* for *geographical* (# 18) and *ethnological* for *ethnic* (# 19). The number of permissible alternatives in a C-Test is usually small, but sometimes decisions about acceptability can be quite tricky, as the *sex hormones/sexual hormones* example shows.

6.3 Wrong answers

On the whole, wrong answers provide us with more insights into text processing strategies than right answers do. As Tables 5 and 6 show, respondents give a variety of different types of responses to the different items.

6.3.1 Incorrect spellings

Several words are spelled wrongly: for instance, *widly* (# 2), *differens* (# 4), *hormons* (# 6, 17), *controll*, *controle* (# 7), *averege* (# 8), *efford* (# 11), *proof* (# 12), *traid* (# 14), *tipical* (# 20), *variaties* (# 21) in HORMONES, *disciplin* (# 2), *showes* (# 10), *linguistik* (# 11), *variabel* (# 15) etc. in

Table 5
Incorrect responses to items on C-Test HORMONES

Experimental work on sex hormones is a politically loaded area, since interpretations of results often end up reinforcing sex role stereotypes.

*It h(1)**as** been*	have
*wid(2)**ely** assumed*	widly, wider*, wide
*th(3)**at** average*	those, this, than*,
*the diffe(4)**rences***	difference, differens, different
*in s(5)**ex***	some, several, sexual
*horm(6)**ones***	hormons
*must contr(7)**ibute***	control, controll, controle, contrast, contract, contradict
*to ave(8)**rage***	averege, avariable, avalon
*differences i(9)**n***	
*behavior. I(10)**n***	It's
*an eff(11)**ort***	efford, effekt, effect
*to pr(12)**ove***	proof, prime
*this, scien(13)**tists***	scientist, science
*have tr(14)**ied***	trying, traid
*to sh(15)**ow***	share
*that varia(16)**tions***	variation§, variability+, variables
*in hor(17)**mone***	hormon, hormonal, hormones, hormones'
*levels wit(18)**hin***	with, without, witch
*each s(19)**ex***	subject, smeed
*are ti(20)**ed***	tidy, tipical
*to varia(21)**tions***	variation§, variability+, variaties, variable
*in beha(22)**viour.***	
*One o(23)**f***	offers, obliges
*the pri(24)**mary***	primaral, primar, prime, private, principal, principle, prior, privilege
*difficulties wi(25)**th***	will

these attempts, however, is that hormone levels themselves are affected by behavior and environment, so that no causal relationships can be established.

Words with the same symbol (*, +, §, $) were used by the same person.

Table 6
Incorrect responses to items on C-Test SOCIOLINGUISTICS

Language is indissolubly linked with the members of the society in which it is spoken, and social factors are inevitably reflected in their speech.

*Sociolinguistics i(1)**s***	in
*he disci(2)**pline***	discission, discition, discision, disciption, disciplin
*which de(3)**als***	demands, demand, decide
*with lang(4)**uage***	languge
*variation cau(5)**sed***	cause, caution, caught, caunt
*by soc(6)**ial***	society*
*differences a(7)**nd***	are, at
*different soc(8)**ial***	society*
*needs. Th(9)**is***	their+
*book sh(10)**ows***	show, showes
*how ling(11)**uists***	lingual, linguistics, linguistic, linging, linguistik
*set ab(12)**out***	abroad, above
*studying i(13)**t.***	is
*It outl(14)**ines***	outly, outline, outlined$, outlays
*the var(15)**ious***	variabel, variation, variable
*social fac(16)**tors***	face, facts
*involved, su(17)**ch***	suggested, succeeded$
*as geogr(18)**aphical***	geographic, geography, geograph
*location, eth(19)**nic***	ethernal, ethic, ether, ethnite, ethial, ethnological, etnic
*origin, soc(20)**ial***	society*
*class, a(21)**nd***	a male, are§, another
*sex, a(22)**nd***	also, all, as, are§
*discusses t(23)**he***	to
*interaction bet(24)**ween***	betwen, betwee
*them. T(25)**he***	their+, these, that, the

book will prove useful to anyone interested in finding out about the complex relationship which exists between language and society.

Words with the same symbol (*, +, §, $) were used by the same person.

SOCIOLINGUISTICS. One of the questions frequently asked is whether wrong spellings should be penalised in the same way as genuinely incorrect responses. From the point of view of test objectivity it is preferable to count as correct only a response which either conforms to the original text or is accepted by native speakers as fitting into the text. Any scorer latitude opens the door to subjectivity. Statistically, it makes very little difference anyway (cf. Grotjahn, Klein-Braley & Raatz, 1992). But the question is, when *is* a word incorrectly spelled? Presumably *hormon*, *hormones* and *hormones'* can all be interpreted as incorrect spelling, but they can equally well be interpreted as genuine misapplication of rules. A large number of otherwise identical words need an additional *e* in English which is not needed in German, so it could be claimed that those subjects who write *hormon* are still working within the German system. In *hormones'* an additional English rule has been misapplied, so this would surely be viewed as a mistake (the greengrocer's apostrophe). How one interprets *proof* (is it a noun where a verb is needed, or merely a misspelling of *prove*?) or *showes* is an open question.

6.3.2 Singular/plural concordance

In a number of cases the problem lies in the use of a plural where a singular is needed and vice versa. Otherwise the word is correct. This is the second frequent type of **narrow focus**, and is demonstrated by *have* (# 1), *difference* (# 4), *scientist*, *science* (# 13), *variation*, *variability* (# 16, 21) in HORMONES, and by *demand* (# 3), *show* (# 10), *outline* (# 14), *these* (# 25) in SOCIOLINGUISTICS. In some cases the focus is very narrow since the word which demands singular/plural concordance is directly adjacent to the damaged item. In other cases there are longer-ranging constraints, for instance in HORMONES, *variations* is separated from the plural verb *are* by six words. Interestingly, if *variation* was offered in # 16 it was repeated in # 21, and the same applies to *variability*. I would hypothesise that the effort of keeping this repetition in mind over the length of the sentence was so great that accurate processing of the verb was impossible, i.e. this is a case of **system overload**.

6.3.3 Right word class, wrong word

Some items show that examinees probably knew which type of word was required, but were not able to find the correct lexical item. For instance in HORMONES, # 7 a verb in the infinitive form is needed, and subjects

produce *control*, *contrast*, *contradict* (correct: *contribute*). In # 24 (correct: *primary*) there is an impressive array of adjective forms offered. *Primaral* and *primar* do not actually exist in English, but show that subjects are trying to use their word-formation knowledge to supply the missing item as do the attempts in SOCIOLINGUISTICS # 19 to produce the adjective *ethnic*. One German subject in this group processing CANALS wrote above the blank for *lock* its German equivalent *Schleuse*, indicating that she knew exactly what was needed, but simply did not have the requisite lexical item available in English.

A phenomenon which belongs in this group, is fairly common, but which does not appear in these particular texts, is use of the wrong preposition, for instance, *on* instead of *of*. Prepositions belong to the most idiomatic parts of the language, and it is very difficult to justify why a specific item collocates with one preposition rather than another (*allergic to*; *allergisch gegen*). Examinees who correctly insert a completion which yields a preposition are producing a higher-level solution than those who do not recognise that one is required and write such things as *it*, *is*, *if*, etc.

Such responses offer evidence that the morphological, syntactical or collocational rule may be available, although the specific lexical filler cannot be found.

6.3.4 Near misses and nonsense words

Sometimes the subject almost knows the right word, but not quite. For instance, it seems to me that *averiable* (# 8, correct: *average*) in HORMONES is an attempt to get at a word which is floating somewhere near the level of consciousness. Similarly, the different responses to SOCIOLINGUISTICS # 2 (correct: *discipline*) and # 11 (correct: *linguists*) show that the subjects have some idea of the word they want, but are not able to home in on it precisely.

Nonsense words are rare in C-Tests – examinees are much more likely simply to ignore the blank, but they do occur. In HORMONES we find *avalon* (# 8, correct: *average*) and *smeed* (# 19, correct: *sex*). If nonsense words can be interpreted as a signal by the test-taker to the tester that he or she feels unfairly treated, then the small number of cases of refusal to take the test seriously indicates that for examinees, at least, the face validity of the C-Test may not be as low as is often feared.

6.3.5 Suppressing/ignoring elements in the text

Not infrequently text processing appears to blank out portions of the text which interfere with the representation of meaning that the test subject is developing. One example here is given in HORMONES # 7 where the solution supplied, *control*, would make sense if the following word **to** is deleted, so that instead of *differences in sex hormones would contribute to differences in behaviour* the text would read *would control differences in* – a version which would also make sense, but which will not fit in the text as given.

7. Statistical analysis of tests

7.1 Statistics for quality control

Normal test quality control analysis involves entering the test data into a file, coding a correct answer as 1 and an incorrect answer as 0. Missing responses can, if desired, be coded separately. Using a suitable computer program (e.g. SPSS), various indices are calculated for each item and for the test as a whole. These include item difficulty indices (P), item discrimination indices (r_{it}), mean scores and standard deviations (s), reliability (r_{tt}) and validity coefficients (r_{tc}).

Normal test quality control procedures are, of course, used in C-Test development. The main deviation from normal procedures lies in the use of the super-item for item analysis (i.e. the scores for whole texts varying in the English tests used here from 0 to 25). Most researchers report Cronbach's alpha-coefficient as an estimate of reliability. In addition, Grotjahn (e.g. 1987a) has used Guttman's λ_2-coefficient. However this is usually higher than alpha (Grotjahn, personal communication), so that alpha, if anything, is underestimating true reliability.

The normal statistical information relevant to test quality control is reported for the C-Tests used in this study in Tables 1 and 2. The discrimination indices for the individual texts in the German group (part-whole correlations, not listed in the table) have a mean value of .70, median .71, and range from .34 to .96. KR-20 reliabilities[4] calculated for each text unit range from .65 to .90, with a mean and median of .80. Alpha reliabilities

[4] Using the individual deletions as items. This is not a procedure to be recommended under normal circumstances.

for C-Tests (5 items) vary between .86 and .91 with a mean of .89 and a median of .88. A variety of validity coefficients have been calculated. The correlations with the DELTA test (r_{DEL}) range from .44 to .93 with a mean value of .71 and a median of .73. A second validity coefficient is provided in the correlations with the Dictation (r_{dict}), which range from .42 to .86 with a mean of .70 and a median of .73. Finally, the correlations between the C-Tests and the total test score (DELTA + DICT, r_{tot}) range from .44 to .93 with a mean value of .73 and a median of .76. For the English group the P-values vary between 53 and 85 with a median of 74. The lowest KR-20 coefficient is .52, the highest .93, with a mean value of .79 and a median of .86. The alpha reliability coefficients for the four tests are .75, .89, .83 and .93. Statistically, these are very satisfactory tests. Since each of the tests was experimental and had not been pretrialled, one would normally not expect to find such high reliability and validity estimates. However, C-Test quality control frequently shows that the tests need no revisions, except quite often in their levels of difficulty. This does not mean that C-Tests do not need checking to ensure that they are performing satisfactorily.

The aim of this study is examine the statistical data of normal test analysis to investigate the possibility of drawing conclusions about linguistic processing.

7.1.1 Item difficulty indices

The mean item difficulty or P shows the percentage of subjects who successfully completed that particular deletion. In a norm-oriented test the ideal overall mean would be around 50%. Individual items are allowed to vary between 20% and 80%, and scores should be approximately normally distributed. Individual blanks in C-Tests vary considerably in difficulty, and there can also be large differences between the difficulty levels of blanks for which the same word represents the correct solution (cf. Klein-Braley, 1985a).

7.1.1.1 Conclusions from low scorers: bottom-up test processing

If the C-Test is a test of general language proficiency, then it should reveal meaningful differences between high and low scorers. The responses of low scorers, while seldom correct, should nevertheless be interpretable in terms of the quality of their language-processing mechanisms. For instance, one would expect that the type of processing demonstrated by the less able

subjects should concentrate on basic grammatical rules and have narrow focus in terms of text content.

Some students do not attempt a text at all – no blanks have any insertion of any kind. For this analysis such scripts were excluded. All other scripts where at least one blank had been attempted were included in the analysis. For the 17 texts of the WS 90/91 no student makes a score of zero. Indeed, the lowest scores are four (five texts), five (two), seven (seven), nine (one), and ten (two). 120 subjects completed text CAR, and seven of these made the lowest score: seven. Table 7 shows the response behaviour of these subjects. Which items are they completing?

*Be particularly careful when buying a used car from a private individual – you have fewer rights than when buying from a trader. Your rig(1)***hts** *will lar(2)***gely** *depend o(3)***n** *what i(4)***s** *said bet(5)***ween** *you a(6)***nd** *the sel(7)***ler** *– that i(8)***s**, *what y(9)***ou** *are to(10)***ld** *about t(11)***he** *condition a(12)***nd** *value o(13)***f** *the c(14)***ar**. *It i(15)***s** *a good id(16)***ea** *to ta(17)***ke** *someone al(18)***ong** *as a wit(19)***ness**. *Better st(20)***ill**, *have t(21)***he** *car insp(22)***ected** *by a(23)***n** *expert. B(24)***ut** *it i(25)***s** *up to you to decide whether you are getting value for money.*

Table 7
Response behaviour of 5 lowest scorers on C-Test CARS

Subject	Item successfully attempted											
1			3			9	15	16	17		24	25
2	1					9	15	16	17	22		25
3						9	15	16	17	22	24	25
4			3		7	9	15	16	17			25
5	1	2	3	4		9		16			24	
Solutions: 1. rights; 2. largely; 3. on; 4. is; 7. seller; 9. you; 15. is; 16. idea; 17. take; 22. inspected; 24. But; 25 is.												

The first point to note is that the correctly solved items range over the whole text from # 1 to # 25. Obviously all subjects (except possibly Subject 3) have at least scanned the **whole** text. The items which are solved by four or more of these subjects include the pronoun *you* (# 9), the verb *is* (twice: # 15, # 25) and the verb *take* (# 17), words very high on the frequency lists. *idea* (# 16) is solved by all subjects, probably because it is part of the set phrase *a good idea*. Three subjects get *on* (# 3) and

but (# 24) right. Two subjects solve *rights* (# 1) – a word which appears unmutilated in the previous sentence – and *inspected* (# 22). Only one subject each solves *largely* (# 2), *is* (# 4) and *seller* (# 7). Note that this *is* forms part of a passive construction, and can thus be assumed to be more difficult (for instance in terms of the number of transformations involved) than #s 15 and 25 which both form part of the set phrase *it is*. The *is* in # 8, which is not solved by any of the weak subjects, forms part of an explanatory aside, which may imply that the difficulty for the weaker students here is on a higher level textually.

In terms of C-Test processing, it would be reasonable to assume that the repetitions of specific lexical items might be among the test items which are easiest for the weakest students. This is not the case. In CAR we have in # 14 a repetition of the word *car*, which is also highly visible in the first or topic sentence. The expectation would be that this is a very easy item. In fact, it has a difficulty level of 84% (the seven items already discussed are all easier). For text processing strategies this implies that lower ability examinees do not necessarily look for higher level text concepts. They work on smaller units. This is, of course, in accordance with the theoretical assumptions, for instance of interactive grammars, or of metacognitive processing.

In addition, the performance of the weaker students can be compared to the performance of the entire group. Those items which five or four of the weak subjects answer correctly are also among the easiest for the whole group: # 9 is solved by 96% of the 120 subjects, # 16 by 93%, # 15 by 97%, # 25 by 79% and # 17 by 99%. Items # 3 and 24 are solved by 95% and 85% respectively. This implies that the parts of the text which test the easiest "rules" – as revealed by the fact that they are easiest for all participants – are those which the weaker students can process successfully. Their behaviour is not random.

Thus, even at a very low level of language proficiency, certain basic grammatical rules can be accessed: even the very weak individuals can pick up some points on the test without apparently understanding very much of what the text is actually about (cf. Grotjahn & Tönshoff, 1992 for a similar result). However these items are not picked up at random – the C-Test is not a guessing game, but a rule-oriented psycholinguistic activity. The fact that those items which are easiest for the entire group are also those which are solved by the weakest students shows that there are response

regularities operating. We may not be able to specify the exact rules – and indeed in terms of the black box theory we need not be able to do so – but these results show that there is a hierarchy of rules and that the easiest rules are most readily solved by the weakest subjects.

It has been suggested that the way less proficient students can "pick up" points reduces the validity of the C-Test: after all, surely only those subjects whose text comprehension reaches an acceptable level should be able to score points on the test. But in fact this ability to tackle easy sections is exactly what we would expect to happen if the C-Test really does function as a measure of general language proficiency: rudimentary rule systems are reflected in C-Test scores which are very low, but significantly different from zero.

7.1.1.2 Rule hierarchies

Can we turn the data round, and set up a predictability matrix in terms of the likelihood of applying certain rules correctly given a specific level of proficiency?

The C-Test text LITERATURE, a much more difficult test item with an almost ideal overall mean score of $P = 52\%$, was processed by 118 students:

> *Many students of society – historians, political scientists, philosophers – find the study of works of literature useful and readily say so. They d(1)***o** *not fe(2)***el** *threatened b(3)***y** *a different ki(4)***nd** *of disci(5)***pline** *or tem(6)***pted** *to ov(7)***er-***stress th(8)***eir** *own subj(9)***ect's** *special myst(10)***eries.** *The hi(11)***gh** *degree o(12)***f** *imagination nece(13)***ssary** *for distin(14)***guished** *work i(15)***n** *the human(16)***ities** *or soc(17)***ial** *sciences ens(18)***ures** *that m(19)***en** *with th(20)***ese** *powers d(21)***o** *not mis(22)***take** *the tech(23)***nical** *boundaries bet(24)***ween** *academic disci(25)***plines** *for divisions within human experience.*

If we arrange the items in terms of rank difficulty, starting with the easiest item to solve, we reach the arrangement of the text shown in Table 8.

This arrangement must be viewed with caution since in our interpretation of the data we have no option but to consider each deletion in isolation. The test subjects, of course, were confronted with the mutilated text, so that the information available to them was different from that available to the analyst. Furthermore we do not know the order in which specific items were processed, so that the amount of information available to the test subject as he or she processed the deletion in question is impossible to determine, but may well be quite different depending on how many of

Table 8
Items of LITERATURE ordered in difficulty rankings

Rank	P	Item #	Text
1	96	4	*by a different* **kind** *of discipline*
2	94	3	*feel threatened* **by** *a different*
3	92	7	*tempted to over-stress their*
4	92	1	*They* **do** *not feel threatened*
5	79	12	*high degree of imagination*
6	77	8	*to over-stress* **their** *own subject's*
7	77	21	*with these powers* **do** *not mistake*
8	76	2	*do not* **feel** *threatened by*
9	71	17	*the humanities or* **social** *sciences*
10	69	15	*work* **in** *the humanities*
11	67	11	*The* **high** *degree of imagination*
12	66	23	*mistake the tech***nical** *boundaries*
13	65	13	*degree of imagination* **necessary** *for*
14	54	24	*boundaries* **between** *academic disciplines*
15	46	5	*different kind of disci***pline** *or tempted*
16	36	20	*men with* **these** *powers do not*
17	33	25	*between academic disci***plines** *for divisions*
18	30	14	*necessary for distin***guished** *work*
19	26	19	*ensures that* **men** *with these powers*
20	22	18	*social sciences* **ens***ures that men*
21	18	10	*own subject's special* **mys***teries*
22	09	22	*do not* **mis***take the technical boundaries*
23	08	9	*their own* **subject's** *special mysteries*
24	05	6	*discipline or* **tempted** *to over-stress*
25	02	16	*work in the* **humanities** *or social*

the other blanks have already been processed. Moreover we know that the blanks are dependent on each other. What is quite clear, however, is that a better performance on the surrounding deletions increases the redundancy of the item being searched for, and makes it easier to solve. Thus there is of necessity an interaction between redundancy in the text and the efficiency of text processing.

Proceeding with all due caution, then, it nevertheless seems to me that there is evidence here for a hierarchy of rules which agrees on the whole with my subjective assessment of their difficulty. The items ranking at positions 1, 5, 10 are parts of set expressions (*a kind of, high degree of, work in*), the item at 2 represents the *by* formation after a passive construction, rank 3 is occupied by the frequent preposition *over* used as a verb prefix ("too much of something"), and supported by identical use of the equivalent prefix in the mother tongue, as is rank 9 (*social sciences*). The items ranked at places 4 and 7 represent verb negation using *do* as a pro-form. # 6 is a possessive pronoun, # 8 the very frequent lexical item *feel* (rank 441 in the Brown Corpus).

The most difficult items are also intuitively reasonable: at rank 25 the concept of *humanities* as a superordinate term for academic disciplines at the same level as *social sciences* may well be unknown. I suspect that *tempted* (rank 24) may be difficult because of failure to sort out that *or* begins a second coordinate clause in the sentence. The apostrophe necessary in # 9 at rank 23 is a further obvious difficulty, particularly since, in general, Saxon genitive is only permissible with animate, human entities.

A similar analysis of the data for the C-Test text JUPITER (mean score $P = 61$) again shows (see Table 9) that a rough assessment of the level of rules operating can be performed on the basis of the solution difficulties of the individual deletions:

> *If we know the distance of a particular satellite from Jupiter and the time of its revolution about that planet, we can compare that time of revolution with the time it would take to revolve about Earth at the same distance from ourselves. The satel(1)***lites** *whip ab(2)***out** *Jupiter mu(3)***ch** *more qui(4)***ckly** *than th(5)***ey** *would ab(6)***out** *the Ea(7)***rth** *and fr(8)***om** *that w(9)***e** *can calc(10)***ulate** *the inte(11)***nsity** *of Jupi(12)***ter's** *gravitational fi(13)***eld** *relative t(14)***o** *Earth's. Sim(15)***ilar** *calculations c(16)***an** *be ma(17)***de** *easily f(18)***or** *any pla(19)***net** *with a sate(20)***llite** *whose dist(21)***ance** *from t(22)***he** *planet a(23)***nd** *period o(24)***f** *revolution c(25)***an** *be determined.*

The items ranked 1, 2, 3, 5, 8, 9 and 11 here all repeat words or structures which have appeared in the unmutilated lead-in sentence. The item ranked in fourth place is the definite article. Ranks 6 and 9 are occupied by the high frequency verb *made*. The most interesting items are *calculate* at rank 7, and *planet* at rank 12. *Calculate* seems as though it should be a

Table 9
Items of LITERATURE ordered in difficulty rankings

Rank	*P*	Item #	Text
1	100	21	*satellite whose dist***ance** *from the planet*
2	100	7	*about the E***arth** *and*
3	86	25	*and period of revolution can be determined*
4	86	22	*whose distance from* **the** *planet*
5	86	20	*any planet with a satel***lite** *whose*
6	86	17	*can be m***ade** *easily for*
7	86	10	*from that we can calc***ulate** *the intensity*
8	81	24	*and period of revolution can be.*
9	81	16	*calculations can be m***ade** *easily*
10	81	9	*from that we can calculate the intensity*
11	81	2	*The satelites whip* **about** *Jupiter*
12	76	19	*easily for any pl***anet** *with a satellite*
13	76	18	*be made easily* **for** *any planet*
14	71	8	*and* **from** *that we can calculate*
15	67	6	*they would* **about** *the Earth*
16	57	14	*gravitational field relative* **to** *Earth's*
17	53	15	*relative to Earth's. Sim***ilar** *calculations*
18	48	23	*from the planet* **and** *period of revolution*
19	48	3	*about Jupiter m***uch** *more quickly*
20	43	1	*The satel***lites** *whip about Jupiter*
21	38	5	*more quickly than* **they** *would about*
22	29	12	*intensity of Jupi***ter's** *gravitational field*
23	19	13	*Jupiter's gravitational* **field**
24	14	4	*much more qu***ickly** *than they*
25	09	11	*the inten***sity** *of Jupiter's*

difficult word, but it is a high level text concept, and is also supported by *calculations* which appears later in the text, which explains why it is easier than would be expected. *Planet*, obviously latent as part of the semantic field, is also reinforced by its unmutilated appearance later in the text. This is picked up by those making acceptable scores.

The items of high difficulty can also be easily explained. *Intensity* (rank 25) and *field* (rank 23) would probably be latent in the semantic field for

the specialist, but not for the non-expert foreigner. At rank 24, *quickly* is an adverb and therefore needs the *-ly* ending. However, even natives frequently leave off the ending in such positions as well, indeed, it could be claimed that adverbs are among the endangered grammatical forms of the English language. For instance, in text RIVERS eight out of the 27 English children who complete it write *heavy travelled river* instead of the correct *heavily travelled*. Germans find the recognition of adverbs in English very difficult since adverbs in German are not morphologically marked. At rank 21 we find another occurrence of an unrecognised Saxon genitive. In particular the finding that the same phenomenon is similar in difficulty in the two texts confirms that the C-Test responses are lining up the "rules" in a psycholinguistically meaningful order.

Similar analysis can be performed for all C-Tests, and in general the rankings of items at the extreme ends of the scales are intuitively convincing. What is not always obvious without close text analysis is why the items in the middle range rank as they do. In JUPITER, for instance, *they*, ranked at 21st position, is difficult to account for unless one assumes that the pro-form with *would*, i.e. a syntactical difficulty, is in some way responsible, which would, of course, extend the concept of "rule" to a text sequence larger than the individual blank.

A question well worth pursuing would be that of whether this type of analysis of C-Tests could be used to predict the performance of individual deletions in new C-Tests. In order to do this a substantial data base of previously administered C-Tests would be required in order to derive quantitative indices for specific rules. Such a data base is not currently available.[5]

7.1.1.3 Are "mistakes" meaningful in terms of C-Test processing?

Perhaps the most interesting item is # 1 in JUPITER whose solution is *satellites*.

> *If we know the distance of a particular satellite from Jupiter and the time of its revolution about that planet, we can compare that time of revolution with the time it would take to revolve about Earth at the same distance from ourselves. The satel(1)***lites** *whip ab(2)***out** *Jupiter ...*

[5] So far as I know, the most extensive data base examined up to the present moment is the one used in this study.

This should be a very easy item since it has appeared in the singular in the unmutilated introductory sentence. In fact, it turns out to be unexpectedly difficult: at 43% more than half the subjects get it wrong. Inspection of the test scripts shows that the most frequent incorrect answer is the singular *satellite*.

How do we interpret this response? Is it merely a slip of the pen? A spelling error? Or a meaningful mistake? I checked the tests themselves to see whether it was a chance finding. Only eight out of 21 subjects (36%) answer correctly. The 13 wrong responses show that 12 subjects (57%) write *satellite*. One respondent leaves the item blank. However, those who get the item **right** have a mean score over the whole test of 18.86; those who get it **wrong** a mean score of only 14.42. This difference is statistically significant at the .002 level ($t = 3.66$, $df = 18$). Moreover item 1 is linked cohesively with # 5 (*they*). Of those who get # 1 right, seven (86%) also get # 5 right. There is a discrepancy in the performance of three (33%) of the twelve subjects with # 1 wrong: they get # 5 right. The others either leave the item blank (4), or write *that* (2), *those*, *this* or *thin*. The cohesive (syntactical) dependency between these items is revealed statistically in the correlation coefficient $\phi = .65$.

Table 10 shows the performance of the two groups of subjects (those who get this item right, and those who get it wrong) on the other tests in the battery. In all cases those with a correct solution for JUPITER # 1 make a **higher** mean score than those who use the incorrect singular form even though not all t-tests are significant at the generally accepted $p \leq .05$ level.[6]

These results are very strong evidence against any theory that the missing *s* is merely a slip of the pen, a careless error or a spelling mistake. Wrong performance on this item is **meaningful** in terms of overall test performance: the missing plural *s* is a genuine indicator of less proficient language processing.

I would attribute this type of apparently careless behaviour to a phenomenon which seems quite common in complex language tasks: the examinees who get this item wrong are suffering from **system overload**. The rule demanding agreement of subject and verb is not applied because

[6] In view of the exploratory nature of this study, multiple t-tests have been carried out without a corresponding adjustment of the level of significance.

Table 10
Test performance on other subtests broken down by performance on JUPITER, item # 1

Subtest	right	wrong	*t*	*df*	probability
DICT	35.00	23.73	2.04	17	.06
GRA1	18.00	15.58	1.66	19	.11
VOC3	31.55	25.91	1.96	19	.07
GRA2	15.77	12.25	1.82	19	.08
GRA3	18.11	14.41	2.08	19	.05
VOC1	17.00	13.91	1.39	19	.18
CTESTCAR	19.56	15.50	1.84	19	.08
CTEST3	19.78	15.17	2.79	19	.01
CTEST4	17.13	12.17	2.96	18	.01
CTESTLIT	15.44	12.42	1.90	19	.07

although the subject **knows** the rule, and would normally **use** it, at the actual moment of processing it is temporarily unavailable since too many other constraints are foregrounded. Theoretically it seems plausible that system overload becomes less of a problem as the internal grammar becomes more efficient. The better students suffer less frequently from apparently unexplainable processing breakdowns, often referred to as "howlers". Incidentally, this phenomenon is easiest to detect in translation into the foreign language.

7.1.1.4 Differences in performance levels between English and German subjects

The German students are on average nineteen years old, and have spent around nine years learning English at school. The English children have an average age of twelve. The English and German groups worked through three common tests: CAR, CANALS and CUISINE.

For the Germans, CAR has an average *P*-value of 71, CANALS 60, CUISINE 75; the English children score 80% on average on CAR, 73% on CANALS and 76% on CUISINE. The first two texts are easier for the English children than for the German students; CUISINE is approximately equal in difficulty. However, we can probably assume that the English children are slightly more proficient than the Germans since CUISINE, which

deals with cultural aspects of the German-speaking countries, is conceptually more accessible to the Germans than to the English children. This shows that text content, however many efforts are made to keep it unmarked, can affect the difficulty level of a text.

Do both groups of respondents deal with the items in the same way? The correlations between the *P*-values of the individual items for the two groups are $r = .48$ (n.s) for CAR, $r = .75$ ($p < .001$) for CANALS, and $r = .53$ ($p < .01$) for CUISINE. This seems to indicate a reasonable degree of similarity in the difficulty levels of the individual items for the two groups.

7.1.1.4.1 Differences in difficulty for German and English groups on C-Test CAR

Table 11 shows the difficulty levels of the individual items for the two groups processing CAR. A number of items are very similar in difficulty for both groups, as can easily be seen from the column **Differences**. Other items, however, show substantial differences. I shall look at the response behaviour for those items where there is a difference greater than 15%. The majority of differences between performance levels show the English children scoring higher than the German students. The items can be grouped into a number of categories. The first category is determined by familiarity (or lack of familiarity) with common language patterns. # 11 shows 96% of English pupils detecting the definite article, while only 78% of the Germans find it. Those Germans who get the item wrong either leave it blank (12%), or offer *trader* (8%). Most of those who get # 11 wrong also fail to solve # 12. Again, the most frequent response is a blank. Other responses are *are*, *as*, *at* and *ascertain*.

The next group of items reveals foreign language learning problems. In # 13 the students are having trouble with prepositions: the overwhelming majority of the Germans who get this one wrong write *on* (18%). Prepositions, of course, belong to the most arbitrary items in any language. # 19 is obviously a vocabulary problem. While virtually all English children know the word *witness* (91%) only 41% of the Germans find it. The most common solution offered is a blank (52%), followed by *with* (4%). # 21 is a contrastive problem. Even advanced German learners of English find the construction *have something done* difficult, and the most frequent response to this item is *to*: they are trying to use the construction *have to inspect the car.*

Table 11
Mean item difficulty of CAR for English and German subjects

Item #	Deletion	*P* English	*P* German	Difference
1	*Your rig(1)***hts**	65	65	00
2	*will lar(2)***gely**	57	55	02
3	*depend o(3)***n**	96	95	01
4	*what i(4)***s**	83	78	05
5	*said bet(5)***ween**	43	66	−23
6	*you a(6)***nd**	57	67	−10
7	*the sel(7)***ler** –	96	86	10
8	*that i(8)***s,**	87	73	14
9	*what y(9)***ou**	87	96	−09
10	*are to(10)***ld**	39	35	04
11	*about t(11)***he**	96	78	18
12	*condition a(12)***nd**	83	68	15
13	*value o(13)***f**	100	78	22
14	*the c(14)***ar.**	87	84	03
15	*It i(15)***s**	100	97	03
16	*a good id(16)***ea**	100	93	07
17	*to ta(17)***ke**	100	99	01
18	*someone al(18)***ong**	52	65	−13
19	*as a wit(19)***ness.**	91	40	51
20	*Better st(20)***ill,**	70	58	12
21	*have t(21)***he**	100	18	82
22	*car insp(22)***ected**	83	75	08
23	*by a(23)***n**	100	60	40
24	*expert. B(24)***ut**	83	85	−02
25	*it i(25)***s**	96	78	18

In five items we find a better performance by the German students than the English children, despite the otherwise better performance on the English group on the text. Of these, only # 5 shows a difference larger than 15%. In # 5 thirteen (54%) of the children write *better*, and seven of them (29%) follow it up in # 6 with *are* (... *better you are the seller* ...) – a narrow focus reinterpretation of the text which would only fit if

Table 12
Mean item difficulty of CANALS
for English and German subjects

Item #	Deletion	*P* English	*P* German	Difference
1	*Many can(1)***als**	75	72	03
2	*then bec(2)***ame**	58	56	02
3	*neglected a(3)***nd**	79	76	03
4	*forgotten un(4)***til**	54	40	14
5	*in fai(5)***rly**	63	24	39
6	*recent ye(6)***ars**	79	68	11
7	*people be(7)***gan**	42	36	06
8	*to rea(8)***lize**	75	88	−13
9	*what th(9)***ey**	79	88	−09
10	*were mis(10)***sing.**	92	80	12
11	*A tr(11)***ip**	46	36	10
12	*or bet(12)***ter**	79	76	03
13	*still a hol(13)***iday**	100	40	60
14	*on a ca(14)***nal**	79	64	15
15	*is fu(15)***ll**	79	88	−09
16	*of inte(16)***rest.**	75	40	35
17	*There i(17)***s**	79	92	−13
18	*the end(18)***less**	38	68	−31
19	*fascination o(19)***f**	79	88	−09
20	*seeing h(20)***ow**	63	48	15
21	*the lo(21)***cks**	29	04	25
22	*work a(22)***nd**	54	56	−02
23	*looking a(23)***t**	58	68	−10
24	*the diff(24)***erent**	92	92	00
25	*kinds o(25)***f** *bridges.*	92	100	−08

the punctuation were changed. The German group also have *better* as the second favourite response (**not** followed by *are* in most cases), but a higher proportion of them get *between*, the correct solution. The English children therefore show a greater preference for reorganizing the text so that it yields a narrow focus meaning. The Germans prefer to leave the deletion blank. Their differing behaviour on this item probably reflects the greater linguistic

maturity of the Germans, and the related ability to deal with more complex language patterns.

7.1.1.4.2 Differences in difficulty for German and English groups on C-Test CANALS

Table 12 shows the equivalent data for CANALS. Here we have a higher intercorrelation of the ranks of item difficulty indices between the groups, and consequently fewer items with large differences in performance.

The failure of the German group to recognise # 5 *fairly* is again connected with the question of common language patterns: most leave the item blank, but one subject writes *faith*, and another *failure* – this could be an associative response to the words *neglected* and *forgotten*. # 13 is another common language pattern: a frequent abbreviation in English for school holidays is *hols*, so it is not surprising that 100% of the English children can restore this word. The majority of Germans leave it blank, but we also find *holder*, *holding* (2) and *hole*.

21 *locks* is a vocabulary question, and it seems many of the Germans have simply not met this word before. A large minority offers *lorries* as a possible solution, which associates with *bridges* later in the text. As I have already mentioned, one subject pencils in the German word, *Schleuse*.

Nine items in this text are solved better by the Germans, but the differences are small except for # 18 *endless*. Both groups leave few blanks, and with three exceptions all offer *ending*. Two English children and one German student write *end of*.

7.1.1.4.3 Interpretation of the differences between English and German responses

From the evidence presented here it seems plausible to claim that to a large extent the same thought processes operate in both groups. However, it is difficult to substantiate this claim since a correct response gives no evidence of how it came about.

There seems to be some evidence that the slightly less proficient performance of the German group on these texts is the result of missing language data (i.e. specific words, common language patterns, prepositional constructions), and it was also possible to detect a contrastive problem. The English group seem to be more innovative in their incorrect responses and to show more variation. They are more likely than the Germans to reorganise the syntax on a local level. When the Germans are at an advantage, as

they are in processing the item *Austria* in CUISINE, they perform better. I have an impression, which results from reading responses from the two groups' tests, but which I cannot at this point back up by exhaustive data, that the Germans are more able to cope with complex language patterns than the English group. Responding correctly to # 5 in CAR (*between*) and # 18 in CANALS (*endless*), both of which are solved better by the Germans, seem to me to involve holding a longer sequence of language in store. The English children are more ready to reinterpret the text, forming shorter pseudo-clauses, even if the resulting sequence is nonsensical. This would back up the theoretical assumptions that increasing maturity gives the speaker greater control over more complex patterns (the Germans) while the English children as natives can call on a much larger data base of language knowledge.

7.1.2 Item discrimination indices

Item discrimination indices are calculated by correlating the response behaviour on the specific item with the overall behaviour on the rest of the items making up the scale. A positive coefficient shows that high scorers tend to get the item right, low scorers to get it wrong. A negative coefficient occurs when the converse applies. Items with negative discrimination indices are discarded in the process of normal test development, since they do not measure the same thing as the rest of the scale. Naturally, in C-Tests (and in cloze tests), discarding negative items would interfere with the deletion pattern of the test and would invalidate the test as a random sample of language processing.

For all tests in the text processing study, item discrimination indices were calculated over the individual deletions. For the seventeen texts completed by German students this meant that a total of 17 x 25 individual coefficients were computed, in all 425 values. Only ten negative coefficients were found (2.35%). In the 16 tests completed by the English children 16×25 values were computed (400) and of these, 26 (6.5%) negative coefficients are found. Thus we have overall more than 95% positive discrimination indices in a total of 35 different test procedures. This is astonishing in view of the fact that all these measures are experimental, that is, they have not been through a cycle of test development.

What does this tell us? The first point to be made is that these figures are at least partly attributable to a statistical artefact: the fact that individual deletions in C-Tests are not repeated independent measurements of the trait. Performance on individual items is related to performance on other items. While in terms of test statistics this is undesirable, in terms of language processing it is a meaningful and satisfying result. Subject response behaviour in processing C-Tests shows that a text is more than a random collection of individual sentences.

The second point is that C-Tests are very satisfactory as tests. Each deletion measures the response behaviour in the same direction as all the other items and the whole scale.

Negative item discrimination indices

What do the negative discrimination indices found in some of the C-Tests mean? First of all, neither of the texts which were completed by large numbers of subjects has any negative coefficient at all (CARS, LITERATURE). Furthermore, there is a tendency for **more** negative coefficients to appear the **smaller** the number of subjects involved in that test. It therefore seems possible that in small sample groups individual behaviour can affect the item discrimination index quite strongly.

Inspection of test scripts shows that negative coefficients occur when two conditions apply:

1. the item is very easy or very difficult;
2. one or more subjects shows unexpected response behaviour, i.e. gets the item right, although the score is otherwise low, or gets the item wrong, although the score is otherwise high.

To investigate this question, I looked at the each of the items affected in the German group's tests. There are two items in BLOOD which have negative discrimination indices: # 8 ($-.09$, $P = 90$) and # 10 ($-.25$, $P = 76$). Those who get # 8 right have a mean test score of 17.11, those who get it wrong, a mean score of 17.33. For # 10 the figures are 16.89 for those who get the item right, and 19.50 for those who get it wrong. Clearly the items are discriminating in the wrong direction.

The relevant text runs:

*nea(7)***rly** *every hu(8)***man** *group exam(9)***ined** *has be(10)***en** *found t(11)***o** *consist o(12)***f** *a mix(13)***ture** *of t(14)***he** *same fo(15)***ur** *blood gro(16)***ups**

Table 13
Response behaviour of individual subjects on BLOOD, items # 8 and # 10

Subject	Score	# 8	# 10
1	23	1	1
2	22	1	1
3	22	1	1
4	21	1	1
5	21	1	1
6	20	1	1
7	20	1	1
8	19	1	1
9	19	1	1
10	19	1	1
11	18	1	1
12	18	0	0
13	17	1	1
14	16	1	1
15	15	1	1
16	15	1	1
17	14	1	1
18	10	1	0
19	10	1	0
20	10	1	0
21	7	0	0
22	4	0	0
Overall mean score	16.36		

Table 13 shows the overall score on the test, and the response behaviour on the items involved. There is a clear ordering of the responses according to the overall score on the test – we have almost perfectly scaled items. Anyone with a score higher than 10 gets # 8 right, anyone with a score higher than 14 gets # 10 right. But subject 12, with a score of 18, fails on both items. If subject 12 is removed from the analysis the item discrimination indices change to .34 for # 8, and to .53 for # 10. Thus it is clear that it is idiosnycratic behaviour from **just one out of 22 subjects** which causes

the negative discrimination index to occur – subject 12's version of the text runs *every hundred group examined has be*_____ .

The same phenomenon can be shown to be operating in all other cases of negative discrimination indices in both the English and the German groups: they are the result of idiosyncratic response behaviour by no more than one or two individuals.

These findings have two implications for C-Test analysis. The first is that the super-item level is clearly the one to be used in test quality control procedures: analysis on the deletion level is too delicate unless the groups are somewhat larger than the ones I have used in this study. In a larger group the behaviour of one individual would not affect the results so drastically since it would be averaged out.

The second is that interpretations of response behaviour to individual items need to be cautious when the groups involved are small. Subject 12 seems to have suffered from a minor breakdown in C-Test processing extending over only two successive deletions – she processes the next five deletions successfully. It would be foolish to base a theory of test processing on the performance of just this one subject. This finding indicates the necessity for great caution in the interpretation of think-aloud protocols.

7.1.3 Inter-item correlations

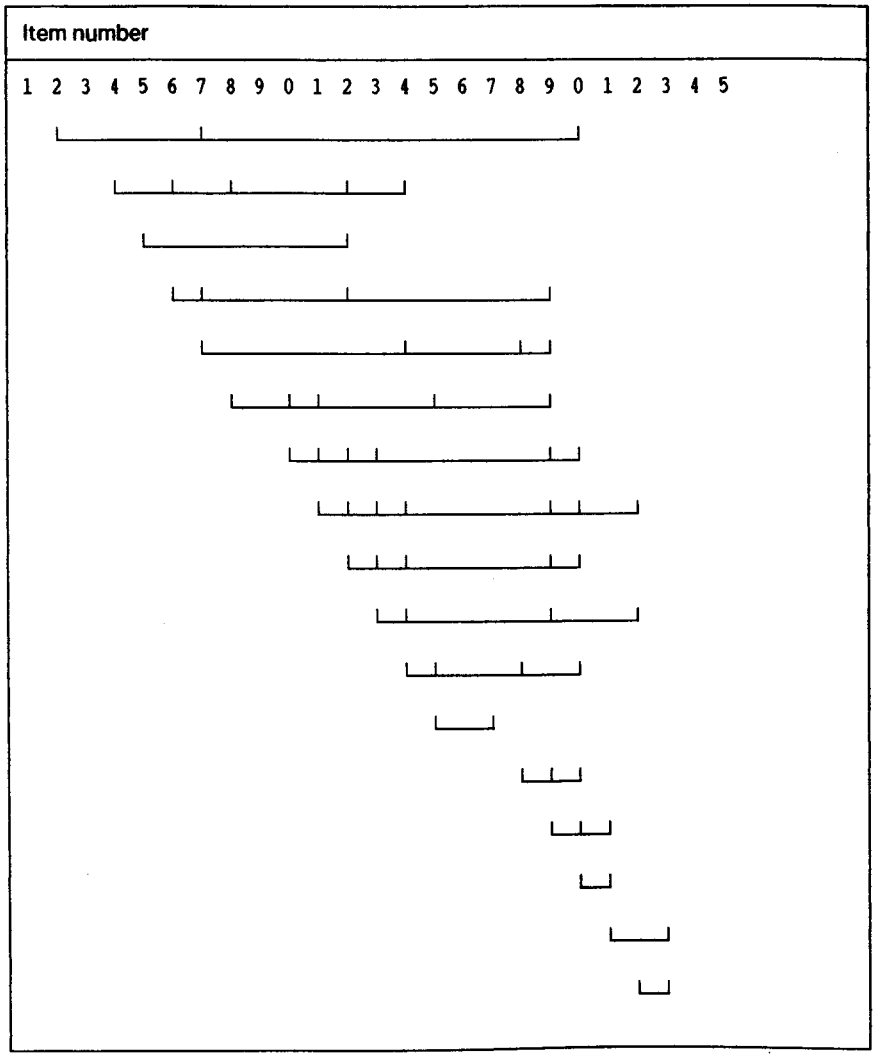
Inside the individual tests the items can correlate positively or negatively with each other. A high positive correlation shows that respondents tend to get both items right – or both items wrong. A negative correlation shows that a correct answer on one item is paired with an incorrect answer on the other. Table 14 shows the positive inter-item correlations higher than .30 for the 25 items of CAR, which was processed by 120 students. Figure 1 shows the same information graphically.

These data show quite clearly that C-Test processing is not merely a question of working one's way from gap to gap. While the pattern in Figure 1 shows that there is a tendency for items in proximity to each other to cluster, for instance #s 10 / 11 / 12 / 13 or 19 / 20 / 21 / 22 / 23, long-range constraints also operate, for instance the correlation of .33 between items 2 and 20, or that of .44 between items 11 and 22. This shows that successful C-Test performance is not merely a matter of lower-level processing; positive correlations between items which are some distance away from each

Table 14
Positive correlations between items on C-Test CAR

1	*Your rig(1)***hts**			
2	*will lar(2)***gely**	7 (.40)	20 (.33)	
3	*depend o(3)***n**			
4	*what i(4)***s**	5 (.45) 12 (.52)	6 (.47) 14 (.42)	8 (.49)
5	*said bet(5)***ween**	6 (.94)	12 (.51)	
6	*you a(6)***nd**	7 (.43)	12 (.53)	19 (.33)
7	*the sel(7)***ler** –	14 (.35) 20 (.34)	18 (.35)	19 (.34)
8	*that i(8)***s,**	10 (.33) 19 (.30)	11 (.33)	15 (.30)
9	*what y(9)***ou**			
10	*are to(10)***ld**	11 (.30) 19 (.33)	12 (.40) 20 (.37)	13 (.39)
11	*about t(11)***he**	12 (.41) 19 (.39)	13 (.46) 20 (.46)	14 (.44) 22 (.44)
12	*condition a(12)***nd**	13 (.59) 20 (.42)	14 (.33)	19 (.39)
13	*value o(13)***f**	14 (.49)	19 (.35)	22 (.35)
14	*the c(14)***ar.**	15 (.30)	18 (.30)	20 (.37)
15	*It i(15)***s**	17 (.45)		
16	*a good id(16)***ea**			
17	*to ta(17)***ke**			
18	*someone al(18)***ong**	19 (.35)	20 (.44)	
19	*as a wit(19)***ness.**	20 (.59)	21 (.30)	
20	*Better st(20)***ill,**	21 (.35)		
21	*have t(21)***he**	23 (.55)		
22	*car insp(22)***ected**	23 (.55)		
23	*by a(23)***n**			
24	*expert. B(24)***ut**			
25	*it i(25)***s**			

Figure 1
Intercorrelations between individual items



other indicate that longer stretches of the text have to be kept in mind for successful performance to take place (cf. also Germann & Grotjahn, 1994).

What are the linguistic implications of this correlation pattern? The cluster at #s 4, 5, 6 and 7 correlates with a second cluster at #s 11 to 14, which in its turn correlates with a third cluster from #s 19 to 23. # 7, which has the highest number of links in the first cluster is the word *seller*. # 14, the centre of the second cluster is the word *car*. # 19, the word with the most links in the third cluster is *witness*. I submit that these three words represent the three highest level text concepts.

Item # 11, *the*, has the most links of any item in the test. It has a difficulty level of 78, astonishingly low for the definite article. Those who get this item right also perform well on #s 12, 13 and 14. This means that subjects either process the phrase *the condition and value of the car* correctly, or they get the whole thing wrong. This cluster links backwards to correct or incorrect performance on the first cluster #s 4 to 7: *what is said between you and the seller*. The long-range correlations forwards relate correct solution of # 11 with another section of the text (#s 19, 20, 22): *... as a witness. Better still, have the car inspected ...* I believe these data show that high level comprehension of C-Tests is necessary for successful test performance.

Negative inter-item correlations

A negative correlation means that one item is processed correctly, the other incorrectly. There are few negative inter-item correlations in CAR, and none is higher than −.2. At the level between −.1 and −.2 the following negative correlations occur:

1 *rights* with # 15 *car* (−.14) and # 25 *is* (−.14);

2 *largely* with # 25 *is* (−.11)

22 *inspected* with # 3 *on* (−.14)

24 *but* with # 16 *idea* (−.11).

These correlation coefficients are probably random – out of 300 coefficients in the matrix 5% would be expected to be negative on chance alone. They largely result from the correlation of easy items, which do not discriminate between subjects, with more difficult items, which do. Naturally, some people who get the easy items right are going to get the more difficult ones wrong. #s 3 (95%), 15 (97%), 16 (93%) and 24 (84%) are all very easy

items, whereas #s 1 (65%), 2 (55%), 22 (75%) and 25 (78%) are more difficult.

7.1.4 Item validity indices

The validity of an item inside a test is its correlation with an criterion outside the scale it is embedded in. In item # 1 in JUPITER apparent carelessness (using the singular rather than the plural form which is demanded by the verb) was shown to be related to performance in all the other tests in the battery. Thorndike (1971) suggests that a validity coefficient of .3 or higher makes a meaningful statement about such a relationship.

7.1.4.1 Overall test performance analysed by response behaviour on individual items

To determine whether the finding for JUPITER was merely a chance occurrence, or whether other individual deletions inside C-Test tests also had the same ability to discriminate well between high and low performance on other tests I investigated the performance of all items on C-Test HEART using performance (a) on the DELTA test and (b) on the combined scores (JOINTC) of CAR and LITERATURE as the criterion.

Tables 15 and 16 show the results. In HEART, #s 11, 12, 19 were answered correctly by all students. No student got # 3 right, so these items are ignored in the analysis. In Table 16 # 5 cannot be analysed because only one student gets the item wrong, scoring 23 points on JOINTC, and a *t*-test cannot be performed.

Table 15 shows that in 19 out of 21 cases, students who got an item right in C-Test HEART achieved a considerably higher mean score on the DELTA test than those who got the item wrong. 10 of the 21 *t*-tests performed over the individual items are significant at the $p < .05$ level. The others do not reach this level of significance, but there are only two instances where the item discriminates in the wrong direction (#s 13 and 15). Similarly, in Table 16, getting the item right results in better JOINTC performance in 19 out of 21 cases. 7 out of 20 *t*-tests are significant at the $p < .05$ level, and only the same two items discriminate in the wrong direction. Item # 13 demands *through*, and # 15 *arteries* (*This blood reaches the muscle through the two coronary arteries*). # 13 has a *P*-value of 87% and # 15 of 52%. Only # 13 has a negative item discrimination index (−.12), although # 15 is very low (.07).

Table 15
Test performance on DELTA broken down by performance on individual items of HEART

Item #	right	wrong	*t*	*df*	probability	correlation
1	131.09	82.83	2.87	21	.009	.53*
2	115.30	43.33	2.90	21	.009	.53*
3						
4	123.56	42.40	5.01	21	.0001	.73**
5	108.95	39.00				.31
6	114.15	51.00	2.43	21	.02	.46
7	125.31	80.70	2.56	21	.01	.48
8	123.93	77.88	2.61	21	.01	.49*
9	106.90	95.50	.33	21	.74	.06
10	127.00	82.91	2.54	21	.02	.48
11						
12						
13	104.30	116.67	−.42	21	.67	−.09
14	129.75	93.20	1.90	21	.07	.39
15	103.17	108.91	−.29	21	.77	−.07
16	111.71	89.50	1.01	21	.32	.25
17	131.67	96.82	1.64	21	.11	.34
18	108.59	98.33	.46	21	.65	.13
19						
20	110.38	89.90	.87	21	.39	.18
21	110.73	96.88	.67	21	.50	.13
22	128.25	81.54	2.75	21	.01	.51*
23	118.71	69.67	2.47	21	.02	.47
24	121.78	48.80	4.06	21	.001	.66**
25	116.05	57.75	2.56	21	.01	.48

* significant at the .05 level
** significant at the .01 level

These findings are extraordinary. They indicate an exceptionally accurate discrimination efficiency on the part of **individual** C-Test blanks in their power to distinguish between different levels of student proficiency in English. The final columns of the two tables, which show the correlation

Table 16
Test performance on JOINTC (CAR + LITERATURE) broken down by performance on individual items of HEART

Item #	right	wrong	*t*	*df*	probability	correlation
1	33.10	25.00	2.91	20	.009	.54*
2	29.84	21.33	1.92	20	.07	.39
3						
4	30.59	22.20	2.42	20	.02	.47
5	28.95	23.00				.16
6	29.63	22.67	1.53	20	.14	.32
7	32.50	24.10	3.07	20	.006	.56*
8	30.31	26.33	1.23	20	.23	.26
9	28.75	28.00	.13	20	.89	.02
10	31.27	26.09	1.67	20	.11	.35
11						
12						
13	28.63	29.00	−.08	20	.94	−.01
14	31.63	27.00	1.41	20	.17	.30
15	27.55	29.81	−.70	20	.49	−.15
16	29.35	26.40	.76	20	.45	.16
17	32.33	27.31	1.42	20	.17	.30
18	29.47	26.00	.90	20	.38	.19
19						
20	29.23	26.80	.62	20	.54	.13
21	30.43	25.62	1.47	20	.16	.31
22	32.18	25.18	2.41	20	.02	.47
23	31.00	22.50	2.67	20	.01	.51*
24	30.35	23.00	2.05	20	.05	.41
25	30.05	22.50	1.92	20	.07	.39

* significant at the .05 level

of the individual item with the criterion, also contain some quite extraordinary figures, for instance the correlation of .73 ($p < .001$) of BLOOD, # 4 *of – Like all muscles it needs a plentiful supply* **of** *blood* – with the overall score on DELTA.

7.1.4.2 Validation coefficients for CANALS processed by the English group

Table 17 presents data for the English children working on C-Test text CANALS. Unfortunately we have no validation criteria other than the C-Tests for this group. There is no item in the test which does not have a correlation of at least .36 with one of the other C-Tests, showing that individual items are measuring the same thing that other entire C-Tests measure, though it seems as if specific aspects of language processing may be picked up at different levels by individual items and C-Tests. This would not detract from the claim that C-Tests measure overall language proficiency, since this holistic entity must consist of subskills and subcompetences, and any C-Test or item can only sample individual aspects of these.

7.1.4.3 Validation coefficients for CANALS processed by the German group

Table 18 shows CANALS as processed by the German group. Here more extensive data are available. In the table I have listed the correlation coefficients between individual items of CANALS, individual DELTA subtests, and the C-Tests CAR and LITERATURE.

I should like to interpret the DELTA tests in the following way: the scores on tests called GRA are connected with the knowledge of grammar as evidenced by the active and passive ability to use traditional grammatical rules as conventionally defined by teachers and textbooks – the rules encapsulated in the individual test items in the GRA tests – to get the items on the DELTA subtests right. Tests called VOC show active and passive command of basic vocabulary. DICT tests the examinee's overall on-line language processing proficiency by presenting her or him with an incoming stream of sound to process which has to be written down according to the orthographical rules of British or American English. C-Tests CAR and LITERATURE are parallel tests to CANAL and test the construct "general language proficiency".

If these assumptions are permitted, then Table 18 shows that items 2, 3, 6, 7, 8, 10, 19, 22, 23 and 25 are linked to grammatical knowledge, producing correlations higher than .5 with at least one of the tests labelled GRA. Items 2, 4, 8, 21 correlate at .5 or higher with a test labelled VOC. Items 2, 8, 22, 23 correlate higher than .5 with DICT and show overall on-line processing proficiency in language. Items 1, 5, 6, 8, 9, 10, 14, 23 correlate at .5 or higher with CAR or LITERATURE. For all items there is

Table 17
Text CANALS (English group)
***P*-values, discrimination indices and correlations with C-Tests PETS, AMERICAN and CORNWALL**

item	*P*	r_{it}	correlations > .5 with other C-Tests ($p < .001$) *(otherwise highest value in italics)*
Much of Britain is criss-crossed by canals that used to be full of heavily laden boats drawn by horses on the towpaths until the railways and then the lorries took over. Many			PETS / AME-RICAN / CORN-WALL
*can(1)***als** *then*	75	.70	(*.44*)
*bec(2)***ame** *neglected*	58	.40	(*.39*)
*a(3)***nd** *forgotten*	79	.67	yes
*un(4)***til** *in*	54	.68	yes / yes / yes
*fai(5)***rly** *recent*	62	.53	yes
*ye(6)***ars** *people*	79	.55	yes
*be(7)***gan** *to*	41	.47	yes / yes
*rea(8)***lize** *what*	75	.46	yes
*th(9)***ey** *were*	79	.43	(*.42*)
*mis(10)***sing**. *A*	91	.35	(*.41*)
*tr(11)***ip** *or*	45	.67	yes
*bet(12)***ter** *still a*	79	.21	(*.36*)
*hol(13)***iday** *on a*	100	–	
*ca(14)***nal** *is*	79	.51	yes
*fu(15)***ll** *of*	79	.71	(*.36*)
*inte(16)***rest**. *There*	75	.66	yes / yes / yes(*.82*!)
*i(17)***s** *the*	79	.69	yes
*end(18)***less** *fascination*	37	.60	yes
*o(19)***f** *seeing*	79	.62	yes
*h(20)***ow** *the*	62	.80	yes / yes
*lo(21)***cks** *work*	29	.58	yes
*a(22)***nd** *looking*	54	.51	yes / yes / yes
*a(23)***t** *the*	58	.52	yes
*diff(24)***erent**	91	.61	yes
*kinds o(25)***f**	91	.61	yes
bridges. Even if you live in a city there may be a canal running through it with lots of interesting sights.			

Table 18
Text CANALS (German group)
***P*-values, discrimination indices and correlations with DELTA subtests and C-Tests CAR and LITERATURE**

item	*P*	r_{it}	correlations > .5 with DELTA subtests ($p < .001$) *(otherwise highest value in italics)*
Much of Britain is criss-crossed by canals that used to be full of heavily laden boats drawn by horses on the towpaths until the railways and then the lorries took over. Many			
*can(1)***als** *then*	72	.67	CAR
*bec(2)***ame** *neglected*	56	.56	DICT, VOC3, GRA2, VOC1
*a(3)***nd** *forgotten*	76	.51	GRA3
*un(4)***til** *in*	40	.35	VOC1
*fai(5)***rly** *recent*	24	.28	LITERATURE
*ye(6)***ars** *people*	68	.62	GRA1, GRA3, CAR
*be(7)***gan** *to*	36	.24	GRA1
*rea(8)***lize** *what*	88	.47	DICT, VOC3, GRA3, CAR
*th(9)***ey** *were*	88	.71	CAR
*mis(10)***sing**. *A*	80	.72	GRA1, CAR
*tr(11)***ip** *or*	36	.41	*(.33, CAR)*
*bet(12)***ter** *still a*	76	.08	*(-.14, GRA1, CAR .37)*
*hol(13)***iday** *on a*	40	.40	*(.43, CAR)*
*ca(14)***nal** *is*	64	.57	CAR
*fu(15)***ll** *of*	88	.60	*(.31, LITERATURE)*
*inte(16)***rest**. *There*	40	.42	*(.27, CAR)*
*i(17)***s** *the*	92	.20	*(.41, LITERATURE)*
*end(18)***less** *fascination*	68	.27	*(.18, LITERATURE)*
*o(19)***f** *seeing*	88	.34	GRA2, LITERATURE
*h(20)***ow** *the*	48	.36	*(.26, GRA1)*
*lo(21)***cks** *work*	04	.29	VOC3
*a(22)***nd** *looking*	56	.46	DICT, GRA1
*a(23)***t** *the*	68	.56	DICT, GRA1, GRA2, GRA3, CAR
*diff(24)***erent**	92	.26	GRA1
*kinds o(25)***f**	100	–	
bridges. Even if you live in a city there may be a canal running through it with lots of interesting sights.			

at least one correlation with another test of at least .18. There is just one figure in the table which does not fit into the overall pattern, namely the negative correlation of −.14 between CANALS # 12 and GRA1. Otherwise the data presented in Table 19 can be interpreted as showing that individual C-Test items are related to grammatical and lexical rules, and to language processing strategies. Many items have multiple relationships. This picture provides a further legitimation for claiming that C-Tests are tests of overall language proficiency.

7.1.5 Factor analysis

7.1.5.1 DELTA subtests and C-Tests CAR and LITERATURE

As Raatz (1982) and others have pointed out, the primary function of factor analysis as an exploratory technique is that of investigating patternings in the data which might not otherwise reveal themselves. Table 19 shows the results of factor analyses calculated over the data for the German group, using the six DELTA subtests: DICT, GRA1 to GRA3, VOC1 and VOC3 and those C-Tests which all subjects completed: CAR and LITERATURE. In analysis 1 all tests load on one factor, which has an eigenvalue of 5.17 and explains 64.7% of the total variance. The two C-Tests have the lowest loadings on this factor.

Table 19
Factor analysis over DELTA subtests and C-Tests CAR and LITERATURE

SUBTEST	FACTOR ANALYSIS 1	FACTOR ANALYSIS 2
DICT	.88	.87
GRA1	.85	.85
VOC3	.85	.85
GRA2	.80	.81
GRA3	.85	.85
VOC1	.87	.87
CAR	.69	**
LITERATURE	.58	**
JOINT C-TESTS	**	.70

Normally, of course, the five texts of the C-Test would run in a factor analysis as one joint unit. Here, if we add the two C-Tests and analyse them as one unit (Table 19, analysis 2), the eigenvalue rises to 5.18 (a genuine increase since there is one test fewer in the analysis), and the total variance explained is 74.1%. The normal pattern with the DELTA data is that all subtests, including the C-Test, have loadings on the general factor somewhere in the region between .7 and .9, even when the C-Test is a new experimental version.

7.1.5.2 Analysis on an item level

Factor analysis on the basis of individual items is a more dubious procedure since the factors which emerge may be difficulty factors rather than factors which are meaningful in some substantial way. However, in an exploratory spirit, the individual items of CAR and LITERATURE were each entered into a factor analysis, together with the DELTA subtests.

It is normal practice to interpret only those factors whose eigenvalue is higher than 1.0. The question of how high a loading has to be in order to be interpretable or meaningful is one which can only be answered heuristically; here I have chosen to use .4 as the critical value.

7.1.5.3 Factor analysis of C-Test CAR

*Be particularly careful when buying a used car from a private individual – you have fewer rights than when buying from a trader. Your rig(1)***hts** *will lar(2)***gely** *depend o(3)***n** *what i(4)***s** *said bet(5)***ween** *you a(6)***nd** *the sel(7)***ler** *– that i(8)***s**, *what y(9)***ou** *are to(10)***ld** *about t(11)***he** *condition a(12)***nd** *value o(13)***f** *the c(14)***ar**. *It i(15)***s** *a good id(16)***ea** *to ta(17)***ke** *someone al(18)***ong** *as a wit(19)***ness**. *Better st(20)***ill**, *have t(21)***he** *car insp(22)***ected** *by a(23)***n** *expert. B(24)***ut** *it i(25)***s** *up to you to decide whether you are getting value for money.*

In the initial solution, the factor analysis for CAR produces eight factors with an eigenvalue higher than 1.0, and four in the rotated solution. At least one item loads with a value higher than .4 on each of the eight factors extracted. The items in CAR with loadings higher than .4 on each of the factors in the rotated solution (Varimax) are shown in Table 20.

Table 20
Results of item factor analysis for C-Test CAR (Varimax rotation)

The first decimal value represents the loading of the item on the factor; the second value in brackets is the mean difficulty P of the item in the C-Test			
Factor 1: Eigenvalue 5.63, explained variance 22.6%			
#11	.57	(78):	*about t(11)***he** *condition*
#12	.45	(68):	*a(12)***nd** *value*
#13	.41	(78):	*o(13)***f** *the c(14)***ar**
#18	.51	(65):	*someone al(18)***ong** *as*
#19	.61	(40):	*a wit(19)***ness.**
#20	.71	(58):	*Better st(20)***ill,**
#21	.48	(18):	*have t(21)***he** *car*
#22	.57	(75):	*insp(22)***ected** *by*
#23	.46	(60):	*a(23)***n** *expert.*
Factor 2: Eigenvalue 1.64, explained variance 6.6%			
#5	.88	(65):	*bet(5)***ween** *you*
#6	.90	(66):	*a(6)***nd** *the seller*
#12	.48	(68):	*a(12)***nd** *value*
Factor 3: Eigenvalue 1.54, explained variance 6.2%			
#4	.46	(78):	*what i(4)***s** *said*
#8	.53	(73):	*that i(8)***s,**
#14	.45	(84):	*o(13)***f** *the c(14)***ar.**
#15	.66	(97):	*It i(15)***s** *a good id(16)***ea**
#17	.52	(99):	*to ta(17)***ke** *someone*
Factor 4: Eigenvalue 1.14, explained variance 4.6%			
#7	.48	(86):	*a(6)***nd** *the sel(7)***ler**
#16	.46	(93):	*a good id(16)***ea**

7.1.5.4 Factor analysis of C-Test LITERATURE

The factor analysis for LITERATURE shows much the same picture as that for CAR.

> *Many students of society – historians, political scientists, philosophers – find the study of works of literature useful and readily say so. They d(1)***o** *not fe(2)***el** *threatened b(3)***y** *a different ki(4)***nd** *of disci(5)***pline** *or tem(6)***pted** *to ov(7)***er-***stress th(8)***eir** *own subj(9)***ect's** *special myst(10)***eries.** *The hi(11)***gh** *degree o(12)***f** *imagination nece(13)***ssary** *for distin(14)***guished** *work i(15)***n** *the human(16)***ities** *or soc(17)***ial** *sciences ens(18)***ures** *that m(19)***en** *with th(20)***ese** *powers d(21)***o** *not mis(22)***take** *the tech(23)***nical** *boundaries bet(24)***ween** *academic disci(25)***plines** *for divisions within human experience.*

Here, the initial solution produces ten factors with eigenvalues higher than 1.0. As Table 21 shows, after rotation four factors remain. The items with loadings higher than .4 on each of the rotated factors are listed in the table.

7.1.5.5 Interpretation of factor analyses in terms of linguistic processing

What conclusions can be drawn from these results? First of all, it seems to me that the items which load on factor 1 for both texts represent segments of the text which are central to its content: *about the value and condition of the car; take someone along as a witness. Better still, have the car inspected by an expert* in CAR; *the high degree of imagination; men with these powers; the technical boundaries between academic* ... in LITERATURE. Factor 1 in both cases therefore represents the content of the text, and shows its importance in solving the C-Test.

Whole running sections of text can appear on the same factor because of the phenomenon that adjacent items show a tendency to load on the same factor. This is more marked in CAR, with one run of three, and one run of six items in sequence. Factors 2 and 4 also show pairs of items. Factors can emerge if the items are merely similar in difficulty, and naturally, in some cases, these adjacent items have very similar P-values – although the fact that they are adjacent items seems to rule out the possibility that they could simply be difficulty factors. Furthermore, other sets of adjacent items loading together on the same factor do not have the same level of difficulty. These findings clearly reveal the interdependencies between items which operate when the items are mutilations performed on words following each other within one single coherent text.

Table 21
Results of item factor analysis for C-test LITERATURE
Varimax rotation

The first decimal value represents the loading of the item on the factor; the second value in brackets is the mean difficulty P of the item in the C-Test			
Factor 1: Eigenvalue 3.50, explained variance 14.0%			
#11	.49	(67):	*The hi(11)**gh** degree*
#12	.42	(79):	*o(12)**f** imagination*
#19	.46	(26):	*that m(19)**en** with*
#20	.47	(36):	*th(20)**ese** powers*
#23	.47	(66):	*the tech(23)**nical** boundaries*
#24	.46	(54):	*bet(24)**ween** academic disci(25)**plines***
Factor 2: Eigenvalue 1.95, explained variance 7.8%			
#1	.43	(92):	*d(1)o not fe(2)**el** threatened*
#17	.47	(71):	*soc(17)**ial** sciences ens(18)**ures** that*
#21	.51	(77):	*d(21)o not mis(22)**take** the tech(23)**nical***
#25	.43	(33):	*boundaries bet(24)**ween** academic disci(25)**plines** ...*
Factor 3: Eigenvalue 1.90, explained variance 7.6%			
#2	.52	(76):	*fe(2)**el** threatened b(3)**y***
#5	.45	(46):	*a different ki(4)**nd** of disci(5)**pline***
Factor 4: Eigenvalue 1.81, explained variance 7.2%			
#13	.48	(65):	*imagination nece(13)**ssary** for*
#18	.56	(23):	*ens(18)**ures** that m(19)**en***

Thirdly, there are three cases where items loading on a single factor involve the same structure: in CAR, factor 2, two items mutilate "***and***", on factor 3 "***is***" is affected three times. In LITERATURE, factor 2, "***do***" is mutilated twice. While the P-values of the first pair are very similar, those of the the second pair differ by 20%. The three "***is***" items have difficulty values ranging from 73 to 97. Here, quite unobtrusively, the C-Test seems to be picking up the ability to use a "rule" correctly or incorrectly. Table 22 shows the behaviour of CAR, #s 4 and 15 (***is***) and LITERATURE #s 1 and 21 (***do not***). In CAR, 80% of the respondents either get both items right, or both items wrong; in LITERATURE, slightly over 80%.

Table 22
Crosstabulations of items where the same structures load on the same factor

CAR # 4 by CAR # 8	0	1	LIT # 1 by LIT # 21	0	1
0	18 (15%)	9 (7.5%)	0	7 (5.9%)	2 (2.2%)
1	14 (11.7%)	79 (65.8%)	1	20 (16.9%)	89 (75.4%)
$\phi = .478$, $p < .0001$ $\hat{\chi}_1^2 = 25.92$, $p < .0001$			$\phi = .375$, $p < .0001$ $\hat{\chi}_1^2 = 13.44$, $p < .002$		

There are no obvious reasons for the groupings of other items on the various factors. Factors 3 and 4 of LITERATURE show loadings for content words (*feel*, *discipline* on factor 3; *necessary*, *ensures* on factor 4), as does factor 3 of CAR (*seller*, *idea*). There may be long-distance dependencies between the response behaviour on these items which are not easy to detect: we are still working, after all, with, and from the outside of, the black box.

8. Top-down text processing

It seems reasonable to assume that as language skills develop, so will the ability to make more use of the more efficient top-down processing strategies (cf. e.g. Lutjeharms, 1988). One of the main problems with C-Test analysis is that although it seems logical that top-down processing must take place

– because otherwise the text could not be satisfactorily reconstructed – it is very difficult to detect in action. This is partly a result of the fact that the texts involved are very short (see Grotjahn, 1996a for a similar argument).

8.1 Correct solutions and crossings out

Respondents who produce correct solutions to mutilated words are presumably producing their responses for the right reasons. Normally, we have no direct way of discovering what these are. However, in some few cases we can follow the examinee's thought processes to a limited extent, namely when someone changes their mind and crosses out the initial solution, replacing it with a second answer.

Tables 23 and 24 show the crossings out which were detected in the two C-Tests FRENCH and WORK. The final solution is the first word given; the word crossed out is given in brackets and marked as crossed out. In no case here is a correct solution falsely rejected, although I have seen instances of this happening.

Analysis of these items shows that more than one cycle of response has taken place. In virtually all cases the solution finally arrived at is the correct solution to the item. In FRENCH there are two items where no solution was finally offered since *somebody* at # 2 (correct: *sometimes*) and *diagonally* at # 18 (correct: *dialect*) do not fit. Obviously the test subjects in these two cases are simply trying out any item which occurs to them on the basis of the letters left standing: here, the C-Test really **is** a guessing game. But the guess is then rejected, which shows textual awareness and at least a start on a second cycle of text processing. In # 7 *centenary*, # 8 *unlimited* and # 20 *pressing*, where incorrect first solutions are replaced by correct versions, we can also assume an initial guessing strategy which leads to a wrong solution which is then replaced in a second cycle of processing, taking more of the text into consideration, by the correct one. How far analysis of the surrounding text, the near and far context, the morphology, the grammar etc. led to substitution of the correct for the incorrect solution cannot be detected.

Table 23
Crossings out in C-Test FRENCH*

Attitudes towards the French language have been deeply influenced by language policies developed in France since the seventeenth century.

*Through vigo(1)***rous**	
*and some(2)***times**	— (~~somebody~~)
*brutal lang(3)***uage**	
*planning progr(4)***ammes,**	
*multilingual Fra(5)***nce**	
*emerged i(6)***n**	
*this cen(7)***tury**	century (~~centenary~~)
*as a unili(8)***ngual**	unilingual (~~unlimited~~)
*French st(9)***ate.**	
*In addi(10)***tion**	
*to legis(11)***lation**	
*against n(12)***on**	non (~~not~~) / non (~~new~~)
*-French langu(13)***ages,**	
*policy mak(14)***ers**	
*in Fra(15)***nce**	
*ensured th(16)***at**	
*only t(17)***he**	the (~~towards~~)
*Paris dia(18)***lect**	— (~~diagonally~~)
*emerged a(19)***s**	as (~~and~~, ~~at~~)
*the pres(20)***ent**	present (~~pressing~~)
*standard fo(21)***rm**	
*of t(22)***he**	the (~~two~~) / the (~~that~~)
*language. B(23)***y**	By (~~But~~)
*virtue o(24)***f**	of (~~or~~) / of (~~or~~) / of (~~on~~)
*its popul(25)***ation**	

and cultural vitality, France today can still be considered the heart of the Francophone world.

* The final solution is the first word given; the word crossed out is given in brackets and marked as crossed out.

Table 24
Crossings out in C-Test WORK*

With the increasing complexity of society work for many people has become more and more simply a means of earning a living. But a rec(1)***ent***

*study o(2)**f***	on (~~of~~)
*the mea(3)**ning***	meaning (~~means~~)
*of wo(4)**rk***	women (~~woman~~) / work (~~women~~) / work (~~working~~)
*among a nati(5)**onal***	national (~~native~~)
sample o(6)f	
*employed m(7)**en***	
*indicates th(8)**at***	that (~~thing~~) / that (~~the~~) / that (~~these~~)
*for mo(9)**st***	most (~~more~~) / most (~~money~~)
*men hav(10)**ing***	having (~~have~~)
*a j(11)**ob***	
*serves ot(12)**her***	others (~~other~~)
*functions th(13)**an***	
*the o(14)**ne***	one (~~only~~) / one (~~ordinary~~) / one (~~ones~~) / one (~~of~~)
*of ear(15)**ning***	
*a liv(16)**ing**.*	living (~~live~~)
*In fa(17)**ct**,*	
*even i(18)**f***	i.e. (~~in~~) / if (~~in~~)
*they h(19)**ave***	had (~~have~~) / had (~~hadn't~~)
*enough mo(20)**ney***	
*to sup(21)**port***	supply (~~supplement~~)
*themselves th(22)**ey***	they (~~that~~)
*would st(23)**ill***	still (~~stop~~, ~~start~~)
*want t(24)**o** work.*	to (~~the~~)
*Wor(25)**king***	

gives them a feeling of having a purpose in life.

* The final solution is the first word given; the word crossed out is given in brackets and marked as crossed out.

8.2 Textual reconstruction: reinterpretation of the text

Textual reconstruction occurs when a test subject reinterprets the text so that it acquires a different meaning. Reinterpretation can operate on short- or long-range segments of the text. Occurrences of this are relatively infrequent, since not all texts lead themselves to reconstruction.

8.2.1 Reprocessing by the German group

In WORK there is an excellent example of this phenomenon in the solution *women* for # 4 (correct: *work*). This yields:

> *a recent study of the meaning of* **women** *among a national sample of employed men indicates that for most men having a job serves other functions than the one of earning a living*

One student who produced *woman* as a first solution was merely unsure whether the singular or the plural was needed, and finally decided on the marginally more correct plural (there is no article present). However another student also produced the solution *women*, and then rejected it in favour of the correct *work*. I think it is reasonable to suggest that this correction can only take place on the basis of understanding at least the entire sentence in which the blank is embedded. This is evidence of higher-level text processing occurring in the filling of a specific blank.

HORMONES

1. *It has been wider assumed than*
   (correct: *it has been widely assumed that)*
2. ***It's** an effort to prove this*
   (correct: *in an effort to prove that)*
3. *variation ... within each sex are tied to variation in behaviour*
   (correct: *variations ... are tied to variations in behaviour)*
4. ***with** each sex are tied to **the** variable in behaviour*
5. *variations ... are tipical to variate behaviour*

SOCIOLINGUISTICS

1. *It outlined the various social factors involved, succeeded ...*
   (correct: *it outlines the various social factors involved such as)*
2. *social differences are different social needs*
   (correct: *social differences and different social needs)*
3. *Their book shows how linguists sit about studying it*
   (correct: *This book shows how linguists set about studying it)*
4. *and a male sex also discusses the interaction*
   (correct: *class, and sex, and discusses the interaction)*

In general, such attempts are unique to one person. Often, punctuation necessary for the new interpretation is inserted. Not infrequently, such text revisions extend over more than one deletion e.g. HORMONES reinterpretation 4 or SOCIOLINGUISTICS reinterpretation 1, where a decision to use a verb with a past tense in # 14 is carried on to # 17 where text parsing breaks down. As this example shows, often the other deletions involved are processed correctly.

Such reinterpretations are obviously the result of not entirely successful text processing, but they do show that longer-range constraints than merely the mutilated word and its immediate neighbours are involved. In SOCIOLINGUISTICS, for instance, we have the word *social* repeated three times, and each time the same respondents replace it with *society*. Or, in the same text, one respondent uses *there* instead of *this* in blanks 9 and 25. It seems to me that such behaviour represents at the very least the retention over a longer period of time of the fact that this particular combination has already appeared once and been "satisfactorily" solved. One could also interpret it as an indicator for high level text processing. There is no way of deciding from the scripts alone.

8.2.2 Reprocessing by the English group

The English children involved in the study engage in more reprocessing. This could be interpreted as a the result of having greater access to alternative units, or of less understanding of the task – not realising that they should recreate the original text as exactly as possible, which is probably a question of maturity. The most likely explanation in my view lies in a combination of the two things.

In C-Test BEHAVIOUR, for the section which reads:

*we(14)***e** *do n(15)***ot** *expect peo(16)***ple** *who d(17)***o** *not bel(18)***ieve** *in G(19)***od** *to g(20)***o** *to chu(21)***rch;**

we find:

1. *We do not expect people who don't not belong in gangs to go to church;*
2. *people who do not belong in Germany to go to church*

and for that which reads:

*We exp(1)***ect** *our fri(2)***ends** *not t(3)***o** *do thi(4)***ngs** *that a(5)***re** *unpleasant f(6)***or** *us;*

we find the following sequence, with the punctuation and the additional *to* added:

3. *We explain our Fridays [,] not today [to] do things that are unpleasant [,] frighten us.*

C-Test text TELEPHONE receives the following reinterpretations:

*Why don't I have a telephone? Not because I pretend to be wise or unusual. There a(1)***re** *two ch(2)***ief** *reasons: bec(3)***ause** *I don't rea(4)***lly** *like t(5)***he** *telephone, a(6)***nd** *because I fi(7)***nd** *I c(8)***an** *still wo(9)***rk** *and pl(10)***ay**, *eat, bre(11)***athe** *and sl(12)***eep** *without i(13)***t**. *Why do(14)***n't** *I like t(15)***he** *telephone? Bec(16)***ause** *I think i(17)***t** *is a pe(18)***st** *and a ti(19)***me***-waster.*

1. *there are two childish reasons*
   there are two choices – reasons [punctuation added]
2. *because I don't really like to telephone anyone*
   because I don't really like to telephone a friend
3. *eat, breakfast and sleep*
   eat, bread and sleep
   eat, bread and slice without it
4. *because I think it is a penny and a time-waster*

The other texts show similar degrees of reprocessing. One impressive example of long-distance reprocessing is demonstrated by the subject who uses *impossible* for both blank 5 and blank 17 in RIVERS:

*The rivers of Germany have long been used as water highways by people. Today i(1)***n** *East a(2)***nd** *West Ger(3)***many**, *water transportation i(4)***s** *still impo(5)***rtant**. *Ships a(6)***re** *permitted t(7)***o** *travel bet(8)***ween**

*these t(9)***wo** *areas b(10)***ut** *they m(11)***ay** *only u(12)***se** *a sm(13)***all** *number o(14)***f** *the conne(15)***cting** *waterways. Germ(16)***any's** *most impo(17)***rtant** *waterway i(18)***s** *the Rhine...*

These data provide evidence that more able subjects do take longer-range constraints into account. Unfortunately, we can identify these processes only in those who make incorrect responses: there is no way to detect how those who complete the text correctly do this. Neither the think-aloud techniques nor this type of analysis enables us to determine precisely the nature of the processing strategies used by the successful subjects.

8.3 Processing under time constraints

Evidence for top-down text processing emerged in a curious way in the present study. Because of an unexpected assembly, the children completing the D-version of the C-Test did not have as much time available as was initially planned, and they were told only to attempt three texts. However anyone who felt that she or he had finished the other tests could go on to the final text. Therefore the results for this group are based on three texts; there are, however, 14 individuals who attempted some or all items in the fourth text.

The first point of interest is that while two of the subjects make very low scores on all four tests (their data have been discarded for this analysis), the scores of twelve of these subjects on the other three texts are virtually perfect. Obviously they have not attempted the fourth text out of desperation or boredom: they feel – rightly – that there is nothing more they can sensibly do on the first three texts. The second point is that with those individuals who began, but did not finish, the fourth text we can look at the blanks they have completed to see where they start text processing.

Table 25 shows the data for CONCRETE as processed by twelve of the high scorers on the D-version of the test.

8.4 Linear or recursive processing?

It is obvious from Table 25 that the majority of pupils do not start at the beginning, go on until they reach the end, and then stop. Subject 12, for instance, begins writing in solutions at # 19, and subject 11, after completing # 1 correctly continues with # 15. The two deletions completed

Table 25
Insertions in C-Test CONCRETE by able English children working under time pressure

With few exceptions, substances expand when heated. If

	Subject number												
	1	2	3	4	5	6	7	8	9	10	11	12	$\sum$
*concrete ro(1)***ad**	·	1	1	1	·	1	1	1	0	1	1	·	8
*surfaces we(2)***re**	1	1	1	·	·	1	1	1	1	1	·	·	8
*laid do(3)***wn**	·	·	1	·	1	1	1	1	1	1	·	·	7
*in o(4)***ne**	·	·	1	·	·	·	1	1	0	0	·	·	3
*continuous pi(5)***ece,**	1	·	0	·	·	·	0	·	·	0	·	·	1
*cracks wo(6)***uld**	1	1	1	·	·	1	1	·	1	1	·	·	7
*appear a(7)***s**	·	·	·	·	·	·	1	·	0	0	·	·	1
*a result o(8)***f**	·	·	1	·	1	·	1	·	·	1	·	·	4
*expansion a(9)***nd** *con-*	·	·	1	·	1	·	1	·	0	1	·	·	4
*traction bro(10)***ught**	·	·	1	·	·	·	1	·	·	1	·	·	3
*about b(11)***y**	·	·	1	·	·	·	1	·	·	1	·	·	3
*the diffe(12)***rences**	1	·	1	·	1	·	1	1	1	1	·	·	7
*between sum(13)***mer**	1	·	1	·	·	·	1	·	1	1	·	·	4
*and win(14)***ter**	1	·	1	·	·	·	1	·	1	1	·	·	5
*temperatures. T(15)***o**	·	1	1	·	1	·	1	·	1	1	1	·	7
*avoid th(16)***is**	·	1	1	·	·	·	1	·	0	1	0	·	4
*the sur(17)***face**	·	1	1	1	·	·	1	·	1	1	1	·	7
*is la(18)***id**	·	1	1	·	1	·	1	·	1	0	1	·	6
*in sm(19)***all**	·	1	1	1	1	·	1	·	·	1	1	0	7
*sections, a(20)***nd**	·	1	1	·	1	·	1	·	1	1	1	·	7
*each o(21)***ne**	·	·	1	·	0	0	1	·	0	0	·	1	3
*is sepa(22)***rated**	1	1	1	1	1	·	1	·	1	1	1	1	10
*from t(23)***he**	1	1	1	1	1	1	1	·	1	1	·	1	10
*next b(24)***y**	·	·	0	0	1	·	1	·	·	0	0	·	2
*a sm(25)***all**	·	·	1	1	1	·	1	0	·	1	0	0	5

gap which is filled with a compound of tar. On a hot summer day this material often squeezes out of the joints on account of the expansion.

1 signifies a correct answer, 0 an incorrect answer, · an item left blank

by virtually all subjects (ten) are # 22 and # 23: *sepa(22)***rated** *from t(23)***he.**

Eight subjects complete the first two blanks: *concrete ro(1)***ad** *surfaces we(2)***re** *laid*. Seven subjects find #s 6 (*would*), 7 (*down*), 12 (*differences*), 15 (*to*), 19 (*small*) and 20 (*and*). This behavour is selective and eclectic, but certainly seems to indicate that subjects scan the text before writing in solutions. They then fill in the blanks where a solution comes to mind immediately.

This recursive method of approaching the text is shown by the wrong solutions offered, since many of them can plausibly be included in the text in a first cycle of processing. For instance, all three wrong solutions to # 5 (*piece*) are *pile*. In #s 19 and 25 the incorrect solutions are *smaller* (correct: *small*). # 19 produces *opening* twice (correct: *one*). Subject 9 offers *october* for # 4 (correct: *one*). Is it coincidence that he is one of those who get the twin items 13 and 14 (*winter and summer*) correct? Other items reveal narrow focus: for # 24 *bit*, *because* (twice), *but* (correct: *by*), or the reversal of *as* and *and* in #s 7 and 9.

The most interesting script is that submitted by subject 10, who writes all correct items and one technically incorrect item in ballpoint pen. The five incorrect items are written in pencil. I take this distinction to mean that she regards the solutions in ballpoint as definitive, the ones in pencil as tentative. All her wrong solutions are plausible in the microcontext, but, as she obviously realises, they do not fit into the macrocontext. Her (technically wrong) solution to # 21 *opening* (correct: *one*) could possibly be accepted as correct.

In pencil she writes *only* for # 4 and *pile* for # 5. She seems to envisage something like *in only [a] continuous pile* – showing a common tendency to read a word that is not actually present. For # 7 she writes *and*, which would yield ... *would appear and* (correct: *as*) *a result of expansion and contraction* .... This seems to be a case of early closure. # 18 emerges as *layered*, which is factually wrong, but definitely plausible in this context (correct: *laid*). Her solution to # 24 *but* (for *by*) could be viewed as an example of something akin to early closure: *but a small gap which is filled with a compound of tar* ... – she seems to assume a sentence continuation which is not present. This subject scores 91 out of a possible 100 points on the whole test, and is obviously one of the most able children. Her idiosyncratic use of different writing instruments enables us to see how she

herself views the solutions. Possibly if she had had more time she would have gone back to the ones which were pencilled in and produced a perfect text, which would have provided no material for interpretation at all.

8.5 Pragmatic text processing

The full set of data presented above in Table 2 for the English group working on CUISINE shows many of the features in subject responses already discussed. For instance, early closure is operating with those respondents who write *low* (correct: *lower*) for # 24, since only the comparative form will enable the sentence to be carried to a satisfactory conclusion. Narrow focus can be seen in # 3, whose correct solution, *so*, is a difficult pro-form. Subjects offer narrow focus solutions such as *seasoning*, *salt*, *salad*, *sauce*, producing phrases like *less salt on fish and vegetables*, which are satisfactory in isolation but fail to fit the syntax of the ongoing sentence.

This text, because of its cultural bias, enables us to look at two phenomena involving a link between the text and the outside world. The first one is the fascinating contrast between *home* and *hotel* in # 11: *wine is generally cheaper than in an English ho*_____. A number of children who originally selected *home* subsequently crossed it out and wrote *hotel*, presumably on the pragmatic assumption that wine is not sold, even in British homes. As can be seen from Table 3 it is the higher scorers who tend to get this item right ($r_{it} = .36$). Fewer of the German group (8 out of 29) chose *home*.

A second item where the differences in response between the English and the German group enable us to see knowledge of the world entering text processing is items 18 and 19 (*Germany and Austria*). None of the German group gets the concept wrong (although two students write *Austrian*), but twelve (50%) of the English group write some form of *Australia*. Obviously the Germans know that Germany and Austria are part of the German-speaking cultural landscape; English children associate *Aus* with *Australia*. In this case there is no connection between getting the item right and making a better score ($r_{it} = .05$). Presumably the difference between the two culturally determined items is that in the first case knowledge of the world has to be connected to logical probabilities; in the second case this is not important.

8.6 Processing a badly formed text

An attempt explicitly to demonstrate pragmatic processing failed to produce any interpretable results. In one of their experiments Bransford & McCarrell (1977) showed that identical, deliberately opaque, texts could be processed more easily if a title which structured them was given to the text. One of their texts was used as a C-Test in a placement test session with a different but equivalent group of subjects:

> The procedure is actually quite simple. First you arrange things into different groups. Of course one pile may be sufficient, depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo things. That is, it is better to do too few things at once than too many. In the short run this may not seem important but complications can easily arise. A mistake can be expensive as well. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life.

The title of this text is WASHING. One group was given the text as a C-Test with its title, the other without. Unfortunately the text was extremely easy in both versions, and, although the subjects provided with the title made higher scores (95%) than those without the title (85%), these differences are not significant.

9. Conclusions

Statistical techniques permit the reduction of large masses of data. Individual statistical indices allow the characterisation of main trends and relationships. Statistical analysis of tests is one technique for determining test quality. Tests should be objective, reliable and valid.

Test validity is the most important characteristic for any test: it should measure what it purports to measure. In the development of the C-Test from the cloze test, Raatz and Klein-Braley (1983) explicitly retained the notion that integrative tests based on the reduction of redundancy were to be viewed as tests of general language proficiency. Statistically, this has never been doubted. High correlations with whatever validation criteria were available show that either **all** the test procedures involved (C-Tests *and* psychometric-structuralist tests, teacher ratings, self-ratings, essays, translations, listening comprehension tests, dictations, oral interviews: cf.

the various articles in Grotjahn, 1992) were measuring meaningful language use – or that **no** procedure was doing so. Therefore, if it is true that C-Tests only tap lower-level processes in language performance this criticism must also be levelled at the other tests and procedures used in validation studies.

Nevertheless the fact remains that the psycholinguist would like to know more about what goes on **psycholinguistically** when subjects process C-Tests. In this study I have used the techniques available in test quality control to relate psycholinguistic and statistical performance. I have shown that responses to C-Tests administered in the real-life testing context of placement can be used to interpret processing operations used by the subjects. I have worked my way through the various levels of text processing, starting with the low scorers picking up scattered points up to the level where it seems legitimate to assume far-ranging text comprehension and pragmatic interpretations at a high level of proficiency. I have shown how the test statistics and the language operations go hand in hand.

I have been able to show that performance on individual items is related to specific operations in language processing. There seem to be meaningful relationships between "rules" as traditionally defined by teachers and linguists and C-Test processing behaviour. Furthermore, the difficulty of "rules" as revealed by arranging individual C-Test deletions in rank order corresponds very well with intuitive assessments of difficulty by the teacher. This relationship between item difficulty levels and language proficiency levels is reinforced by the fact that it is possible to predict the empirical difficulty level of a text for a specific group with a high degree of accuracy (Klein-Braley, 1985b; in preparation).

Some quite extraordinary findings have emerged in the course of the analysis, in particular the relationships of individual C-Test blanks to overall test performance, both on other C-Tests (not so surprising), and on other, more narrow-based, psychometric-structuralist tests of specific phenomena in language.

I have used a large number of texts in order to give the reader an impression of the types of systematic phenomena which can regularly be discovered in C-Test scripts if one once begins to look at them in this way. Any individual C-Test text will reveal some interesting phenomena. But the texts are very short, and the amount of data which can be milked from such a short text is limited. I have tried to counteract this problem by using a variety of texts. Even so, not all the texts listed in Tables 1 and 2 have

been discussed. I can only assure the reader that the texts not discussed here reveal similar processes.

The use of an English and a German group enabled comparisons to be made between the solutions offered and the processing techniques used by native and foreign learners. The results show that in many cases the English group relied more on narrow focus, early closure and text reinterpretation than the Germans. This seems to reflect a less developed ability to process longer stretches of the text on the part of these less proficient subjects, but it also shows the availability of a larger repertoire of responses. The English children also perform better on conventional formulations (set expressions and phrases) and rhetorical devices such as repetition.

There may be a danger inherent in this approach of over-interpreting the data. From responses given by test subjects I have attempted to reconstruct processes possibly operating in item solution. But I cannot prove that these processes take place, although I myself am convinced that this is the case. So far as space permits, I have attempted to provide all the information necessary for readers to follow my argument and make up their own minds. I believe that the C-Test processing mechanisms posited here show that the C-Test taps all levels of text processing. The phenomena we can detect in such a non-reactive approach necessarily operate on the level of the individual blank, but this is not a valid argument for claiming that only low level, local operations are necessary for satisfactory C-Test performance. The data presented here show that long-range constraints do operate, and that successful processing involves keeping these higher-level considerations in mind while simultaneously filling in blanks on the lower levels. System overload, a phenomenon which I first met in the context of translation teaching (cf. Smith & Klein-Braley, 1985, Chapter 12), explains the sudden lapses or howlers which unexpectedly affect students who, if specifically tested, would both "know" the rule and be able to use it. But there are times when complete control of the situation requires more processing capacity than is available. In the context of C-Tests the high scorer is the one who can juggle successfully with the elements on several levels simultaneously; the low scorer suffers at some point from system overload, which leads to breakdown somewhere and to errors in text reconstruction. Or, to restate the thesis: low scorers have less efficient, i.e. less highly developed, processing systems than high scorers. A less highly developed internalised grammar indicates a less advanced level of learning. C-Tests

are able to reveal these differences, and, in view of the simplicity of the test procedures involved, they do it with an extraordinary economy and accuracy.

I believe that the data presented here show clearly that C-Test performance can validly be interpreted in terms of general or overall language proficiency.

References

Bransford, John D. & McCarrel, Nancy S. (1977). A sketch of a cognitive approach to comprehension: some thoughts about understanding what it means to comprehend. In Philip N. Johnson-Laird & Peter C. Wason (Eds.), *Thinking. Readings in cognitive science* (pp. 377-399). Cambridge: Cambridge University Press.

Cohen, Andrew D., Segal, Michal & Weiss Bar-Siman-Tov, Ronit. (1985). The C-Test in Hebrew. In Klein-Braley & Raatz (1985), 121-127.

Feldmann, Ute, Grotjahn, Rüdiger & Stemmer, Brigitte. (1986). Was messen Sprachtests eigentlich? Überlegungen zur introspektiven Validierung von C-Tests. In Seminar für Sprachlehrforschung der Ruhr-Universität Bochum (Ed.), *Probleme und Perspektiven der Sprachlehrforschung. Bochumer Beiträge zum Fremdsprachenunterricht in Forschung und Lehre* (pp. 325-338). Frankfurt/M.: Scriptor.

Feldmann, Ute & Stemmer, Brigitte. (1987). Thin_____ aloud a_____ retrospective da_____ in C-te_____ taking: diffe_____ languages – diff_____ learners – sa_____ approaches? In Claus Færch & Gabriele Kasper (Eds.), *Introspection in second language research* (pp. 251-267). Clevedon: Multilingual Matters.

Germann, Ulrich & Grotjahn, Rüdiger. (1996). Das Lösen von C-Tests auf dem Computer. Eine Pilotuntersuchung zu den Bearbeitungsprozessen. In Grotjahn (1996b).

Grotjahn, Rüdiger. (1986a). Der Bochumer Einstufungstest 'Französisch'. In Seminar für Sprachlehrforschung der Ruhr-Universität Bochum (Ed.), *Probleme und Perspektiven der Sprachlehrforschung. Bochumer Beiträge zum Fremdsprachenunterricht in Forschung und Lehre* (pp. 313-324). Frankfurt/M.: Scriptor.

Grotjahn, Rüdiger. (1986b). Test validation and cognitive psychology: some methodological considerations. *Language Testing, 3*, 159-185.

Grotjahn, Rüdiger. (1987a). How to construct and evaluate a C-Test: A discussion of some problems and some statistical analyses. In Rüdiger Grotjahn, Christine Klein-Braley & Douglas K. Stevenson (Eds.), *Taking their measure: The validity and validation of language tests* (pp. 219-253). Bochum: Brockmeyer.

Grotjahn, Rüdiger. (1987b). On the methodological basis of introspective measures. In Claus Færch & Gabriele Kasper (Eds.), *Introspection in second language research* (pp. 54-81). Clevedon: Multilingual Matters.

Grotjahn, Rüdiger. (Ed.). (1992). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 1). Bochum: Brockmeyer.

Grotjahn, Rüdiger. (1996a). 'Scrambled' C-Tests: Untersuchungen zum Zusammenhang zwischen Lösungsgüte und sequentieller Textstruktur. In Grotjahn (1996b).

Grotjahn, Rüdiger. (Ed.). (1996b). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 3). Bochum: Brockmeyer.

Grotjahn, Rüdiger & Stemmer, Brigitte. (1985). On the development and evaluation of a C-Test for French. In Klein-Braley & Raatz (1985), 101-120.

Grotjahn, Rüdiger, Klein-Braley, Christine & Raatz, Ulrich. (1992). C-Tests in der praktischen Anwending. Erfahrungen beim Bundeswettbewerb Fremdsprachen. In Grotjahn (1992), 263-296.

Hopkins, Edwin A. (1985). Redundancy, entropy and the comprehension difficulty of translation texts: groundwork for the investigation of the strategies of adult German learners of English. In Klein-Braley & Raatz (1985), 132-146.

Klein-Braley, Christine. (1985a). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing, 2*, 76-104.

Klein-Braley, Christine. (1985b). Advance prediction of test difficulty. In Klein-Braley & Raatz (1985), 23-41.

Klein-Braley, Christine. (1985c). C-Tests and construct validity. In Klein-Braley & Raatz (1985), 55-65.

Klein-Braley, Christine. (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.* Habilitationsschrift, Universität Duisburg.

Klein-Braley, Christine & Raatz, Ulrich. (Eds.). (1985). *Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis.* Bochum: AKS.

Little, David & Singleton, David. (1992). The C-Test as an elicitation instrument in second language research. In Grotjahn (1992), 173-192.

Lutjeharms, Madeline. (1988). *Lesen in der Fremdsprache. Versuch einer psycholinguistischen Deutung am Beispiel Deutsch als Fremdsprache.* Bochum: AKS.

Lütticken, Martha. (1985). C-Tests in Spanischkursen an der Volkshochschule. In Klein-Braley & Raatz (1985), 130-131.

Raatz, Ulrich. (1982). Language theory and factor analysis. In Madeline Lutjeharms & Terry Culhane (Eds.). *Practice and problems in language testing 3* (pp. 30-56). Brussel: Vrije Universiteit.

Raatz, Ulrich. (1984). The factorial validity of C-Tests. In Terry Culhane, Christine Klein-Braley & Douglas K. Stevenson (Eds.), *Practice and problems in language testing 7* (pp. 124-139). Colchester: University of Essex.

Raatz, Ulrich. (1985a). Better theory for better tests? *Language Testing, 2*, 60-75.

Raatz, Ulrich. (1985b). C-Tests im muttersprachlichen Unterricht. In Klein-Braley & Raatz (1985), 66-71.

Raatz, Ulrich. (1985c). Investigating the dimensionality of language tests – a new solution to an old problem. In Viljo Kohonen, Hilkka von Essen & Christine Klein-Braley (Eds.), *Practice and problems in language testing 8* (pp. 123-136). Tampere: AFinLa.

Raatz, Ulrich. (1985d). The factorial validity of C-Tests. In Klein-Braley & Raatz (1985), 42-54.

Raatz, Ulrich & Klein-Braley, Christine. (1983). Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis. In Ralf Horn, Karlheinz Ingenkamp & Reinhold Jäger (Eds.), *Tests und Trends 1983, Jahrbuch der Pädagogischen Diagnostik* (pp. 107-138). Weinheim: Beltz.

Raatz, Ulrich & Klein-Braley, Christine. (1992). *Beiheft zu CT-D 4. Schulleistungstest Deutsch für 4. Klassen.* Beltz: Weinheim.

Rumpel, Dieter. (1985). Modelle der Erkennbarkeit von Abkürzungen und ihre tentative Anwendung auf den C-Test. In Klein-Braley & Raatz (1985), 147-160.

Smith, Veronica & Klein-Braley, Christine. (1993). *in other words. Lehrbuch Übersetzung* (3. Aufl.). München: Hueber.
Stemmer, Brigitte. (1991). *What's on a C-test taker's mind: Mental processes in C-test taking*. Bochum: Brockmeyer.
Stemmer, Brigitte. (1992). An alternative approach to C-Test validation. In Grotjahn (1992), 97-144.
Süßmilch, Edgar. (1984). Sprachleistungsmessung mittels C-Tests. *Finlance, 3*, 55-93.
Süßmilch, Edgar. (1985). Tests für ausländische Schüler: Sprachdiagnose im Unterricht Deutsch als Zweitsprache. In Klein-Braley & Raatz (1985), 72-82.
Thorndike, Robert L. (1971). *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education.

Grotjahn, Rüdiger. (Hrsg.). (1996). Der C-Test. Theoretische Grundlagen und praktische Anwendungen (Bd. 3, S. 95–125). Bochum: Brockmeyer

Rüdiger Grotjahn

'Scrambled' C-Tests: Untersuchungen zum Zusammenhang zwischen Lösungsgüte und sequentieller Textstruktur

Störungen in der Konnektivität textueller Strukturen erschweren normalerweise die Verarbeitung eines Textes. Unter der Voraussetzung, daß die Lösung eines C-Tests nicht nur auf der Ebene des Mikrokontextes erfolgt, ist deshalb zu erwarten, daß Änderungen der Reihenfolge von Teilsätzen eines C-Tests zu Verarbeitungsproblemen und damit zu einer Erhöhung der Schwierigkeit sowohl einzelner Items als auch des Gesamttests führen.

Zur experimentellen Überprüfung dieser Hypothese wurden die Teilsätze von drei französischen C-Test-Texten einmal leicht und einmal stärker permutiert (*scrambling*). Zwei Texte waren mit 56 bzw. 52 Lücken länger als üblich; der dritte Text wies dagegen eine fast übliche Länge auf (28 Lücken). Die zwei Permutations-Varianten wurden zusammen mit dem Ausgangstext und dem "Bochumer Diagnostiktest Französisch" mehr als 200 Französischlernern zur Bearbeitung vorgelegt (per Zufallszuweisung).

Die durchgeführten multivariaten statistischen Analysen scheinen die Hypothese zu bestätigen, daß eine Permutation der Teilsätze zu einer Erhöhung der Verarbeitungsschwierigkeit und damit zu einem niedrigeren C-Test-Punktwert führt – jedoch nur bei einem längeren C-Test-Text mit deutlich linearer Struktur. Außerdem scheint es sich um einen sehr schwachen Effekt zu handeln. Insgesamt gesehen scheint die Untersuchung die Auffassung zu bestätigen, daß C-Test-Texte üblicher Länge vor allem auf der Ebene des Mikrokontextes messen. Allerdings ist die vorliegende Studie insbesondere wegen einer nicht optimalen Versuchsplanung nur als eine erste Pilotuntersuchung zu der angesprochenen Thematik anzusehen.

1. Einleitung

Eine Reihe von Autoren (z.B. Cohen, Segal & Weiss Bar-Siman-Tov, 1985; Germann & Grotjahn, 1994; Grotjahn, 1987; Grotjahn & Stemmer, 1996; Grotjahn & Tönshoff, 1992; Kamimoto, 1992; Klein-Braley, 1994, 1996; Stemmer, 1991, 1992) hat sich analog zu den entsprechenden Arbeiten im Bereich der Cloze-Test-Forschung mit der Frage beschäftigt, ob der C-Test über die Ebene des Mikrokontextes hinaus auch die makrokontextuelle Verarbeitung und damit höhere Verstehensprozesse erfaßt. Aufgrund empirischer und theoretischer Befunde kommt die Mehrzahl der Autoren zu dem Schluß, daß der C-Test – zumindest in seiner kanonischen Form mit 20 bis 25 Lücken pro Text – vor allem auf der Ebene des Mikrokontextes mißt