Research in Science Education, 1977, 7, 9-24.

LOGICAL CONNECTIVES IN SCIENCE : A SUMMARY OF THE FINDINGS

P. L. Gardner

Hydrogen stands alone among the elements of the periodic system. Its single electron might make feasible its classification with the alkali metals but hydrogen does not lose its electron easily enough to warrant such classification. Similarly, hydrogen has at times been classified with the halogen elements because it needs only one more electron to satisfy the rare-gas electron arrangement of helium. In any case, trying to classify hydrogen as an alkali metal or as a halogen would only be "forcing" such a classification; for this reason, hydrogen will be considered unique among the elements, constituting a "family" in itself.

Nechamkin (1968, p.1)

This passage, the opening paragraph of the first chapter of a chemistry text, is a fairly typical sample of scientific writing. The paragraph is composed of a number of related propositions. These propositions, and the concepts out of which they are constructed, are linked together to form a coherent argument.

Linkage in English is achieved in many different ways. For example, one form of linkage, known as hyponymy (class-inclusion), is demonstrated in Sentences 1 and 2 of the quoted passage: the 'alkali metals' of Sentence 2 are sub-class of the 'elements of the periodic system' mentioned in Sentence 1. A second form of linkage is brought about by the many function words (e.g. *'among', 'its', 'with', 'such'*) which tie concepts together within a sentence, or carry meaning from one sentence to another.

A third form of linkage is achieved by logical connectives. These are words and phrases which serve as links between sentences, or between a clause and either a phrase or another clause within a sentence. Examples in the quoted passage are *'but', 'similarly', 'because', 'in any case',* and *'for this reason'*. It is this kind of linkage which is the subject of the present research. The Logical Connectives in Science Project has sought to identify those logical connectives which occur most frequently in scientific material, and to investigate secondary school students' difficulties with these terms. The method of compilation of the list of logical connectives, the types of test items employed to investigate difficulties, and some of the preliminary findings from the pilot stages of the project, have been described in two earlier papers (Gardner, 1975, 1976). During 1976, the final versions of the tests were prepared and administered, and the results analysed. A summary of this final phase of the project is given in the present paper.

Testing and sampling

The tests

The final set of tests consisted of 25 different tests, each containing 40 multiple-choice items. Amongst the thousand items in the item bank, half were gap-filling (GF) items, and nearly all the rest were sentence-completion (SC) items. In about half the items, the logical connective was tested in a scientific (S) context; in the remainder it was tested in an everyday (E) context. The following four items testing the work 'actually' are presented for the purposes of illustration:

1. attitudes of enquiry — the *habit* of critical thinking;
2. knowledge of the nature of situations in which the weight or accuracy of statements is to be logically determined;
3. mental skills in applying attitudes and knowledge.

It is the second aspect which lends support to the suggestion that different critical thinking tests are required for different subject areas, for it can be argued that many science situations are different from non-science ones, since they contain aspects of empirical enquiry not common to other areas of knowledge.

Development of the initial test

It was clear that, in order to test aspect (3) using a pencil and paper instrument, I could not expect to improve upon the well-tried approach of the WGCTA. Aspect (1) cannot genuinely be assessed in this manner, but I have proceeded on the assumption that a composite of aspects (2) and (3), which might be termed *cognitive critical thinking ability*, can be validly tested by providing statements referring to material characteristic of the area of study, in this case science, and requiring assessment of these statements by means of evidence and logic. Part of the evidence may be contained within the test material, but part may be required to be contributed by the respondent from his general knowledge, and understanding of the nature of the science subject. General knowledge is a requirement for some items in the WGCTA. The WGCTA consists of five subtests of different format designed to tap somewhat different aspects of the skills required. This subtest structure was retained in the science-based tests used in the work described in this paper.

Since chemistry students were the subject of the study, the first test constructed contained material dealing mainly with chemistry situations. This test was initially part of a battery of tests administered to groups of students, and teachers were not able to make available the double class period necessary for a test of the length of the WGCTA, which has 100 items. For this reason a test of 42 items was produced; it was intended to take 30-35 minutes and will be referred to as Form C.

Examples of WGCTA items, with corresponding chemistry-based items and items from a subsequent physics version (Form P) to be discussed below, are given in Table 1.

Experimental results

For a test in which most of the items are essentially of the true-false type, reliability can be a problem, and in early trials in 1975 this was the case. The test was developed for curriculum evaluation purposes, not for individual assessment, so the low reliability, which in any case could clearly become respectable if the number of items were increased, was not of major importance. However, it was decided for several reasons to produce an alternate form whose only major difference from the first was that most situations were related to physics. Both tests were administered, one week apart in early 1976, to a group of students and the results were used to estimate alternate-forms reliability. Internal-consistency and product-moment correlation data are shown in Table 2.

Form C was administered first, and some students did not finish the final subtest. If this is taken into account, the means are almost identical. This has subsequently remained

TABLE 1

Examples of test items

WGCTA, Form Ym	Recognition of Assumptions	*Stem:* "A wise man will save at least twelve dollars each week out of his earnings". *Proposed assumption:* (is this made in the above statement?): "No fools have sense enough to save twelve dollars a week".
Form P	Recognition of Assumptions	*Stem:* "The continued burning of fossil fuels will cause a catastrophic rise in the earth's temperature due to the insulating effect of increased concentration of atmospheric carbon dioxide". *Proposed assumption:* "Carbon dioxide produced by burning fossil fuels cannot be prevented from entering the atmosphere".
Form Ym	Interpretation	*Stem:* "A report of the U.S. Census during a certain year states that there were approximately 1,656,000 marriages and 264,000 divorces granted in the United States". *Conclusion:* (does this follow from the above?) "The divorce rate in the U.S. is much too high".
Form C	Interpretation	*Stem:* "A chemist, in trying to make two substances react together, used every likely method but to no avail. He gave up, and used the beaker containing the last of the mixture as a receptacle for coffee drugs. The reaction immediately occurred and he obtained a good yield of the required product". *Conclusion:* "The caffeine in the coffee was a catalyst for the reaction".

TABLE 2

Initial results

Form	Mean (/42)	S.D.	Pearson	Reliability (α)
C	26.1	4.89	0.28	0.64
P	27.1	4.67		0.67

N = 45

true, whichever test has been taken first. For such short tests, reliability coefficients are reasonable, but even if uncertainty due to small student numbers is borne in mind, the inter-test correlation can only be described as low.

Investigation showed that all but one or two students were studying both physics and chemistry. The items had been "vetted" and pronounced apparently satisfactory by a number of science educators, and most discriminated satisfactorily in practice. No comparison with correlation data between Forms Ym and Zm of the WGCTA could be made – this important aspect is not covered by the mass of data contained in the test manual.

Two further experiments were carried out in 1976. In the first, each of the three tests was split into two parts, one containing three subtests and the other two. To each part of Form C was added the corresponding parts of Form P and Form Ym, giving two tests of 85 and 99 items. In this way it was hoped to determine the degree of correlation between science and non-science items of the same type.

The figures are subject to uncertainty due to the small number of available students, but some points of interest are contained in the correlation results in Table 3.

TABLE 3

Critical Thinking Sub-Tests. Watson-Glaser Comparisons

C = Form C   P = Form P   W = Watson-Glaser

R = Inference sub-test
A = Assumptions "
D = Deduction "
I = Interpretation "
E = Evaluation "

	WR	PR
CR	0.45	0.34
WR		0.22

	WA	PA
CA	0.56	0.01
WA		0.42

	WD	PD
CD	0.39	0.16
WD		0.48

	WI	PI
CI	0.69	0.74
WI		0.45

N = 23

	WE	PE
CE	-0.15	-0.20
WE		0.40

Indications are that some subtests can more easily than others be made to give fairly consistent results. Thus interpretation (I) items are markedly consistent, but evaluation items much less so. Here the chemistry version is at odds with the others, and it can be seen that in 3 subtests out of 5, the worst correlation is between chemistry and physics items. Again practically all students had studied chemistry and physics.

The second 1976 experiment served to trial revised items for Form P and to provide more data for a comparison between Form C and P. Data were now available from 120 students.

As part of the major evaluation project mentioned earlier, data on Form C were available for a total of 243 cases. In addition, scores on the public examinations in chemistry and physics and on tests of chemistry knowledge, awareness of chemistry-society issues, problem-solving and the nature of science (a shortened version of TOUS) were available for correlation determinations. These are shown in a later table.

For Forms C and P, whole- and sub-test reliabilities and test and subtest intercorrelations are shown in Table 4.

TABLE 4

Combined data

	CSC	PSC	CR	CA	CD	CI	CE	PR	PA	PD	PI	PE
CSC (39)	(N=243) 0.67	.36	.54	.53	.63	.63	.58					
PSC (33)		.67						.72	.56	.76	.70	.75
CR (7)			.41									
CA (7)				.41								
CD (8)					.37							
CI (9)				Mean 0.41	=	.39						
CE (8)							.45					
PR (5)			.15					.31				
PA (5)				.13					.27			
PD (7)			Mean 0.11	=	.04			Mean	=	.34		
PI (5)						.14			0.39		.53	
PE (8)							.09					.36

N = 11

CSC, PSC = Total Scores ( ) = No. of items in subtest

Some items in each tests were judged unsatisfactory and are omitted from all scores.

The major diagonal shows reliabilities (α), the other figures being Pearson correlations. For such small numbers of items, Form C subtests show a non-negligible degree of internal consistency — the inter-subtest correlations are much lower. However, it is doubtful if, for example, a "Test of Evaluation of Arguments" would be useful by itself if based on this format. Watson and Glaser do not recommend using subtest scores because of low reliability coefficients (0.40-0.55 for Form Zm). Clearly, the five types of items devised by Watson and Glaser do not in fact measure distinct skills, nor has it been claimed that they should. Rust (1960) and Follmann (1969) have shown by factor analysis that the ability to do critical thinking tests is a composite of several skills.

The correlation between scores on Forms C and P is still low, and is comparable with the examination correlations shown later in Table 6. The inter-subtest correlations are particularly low. Clearly, knowledge of subject matter is not a major factor. Does the low correlation therefore imply that critical thinking in physics is different from critical thinking in chemistry?

As stated earlier, the tests were not initially intended to be strictly chemistry- or physics-based. In fact, there were two items on Form C which dealt with physics and one on Form P which dealt with chemistry. A number of items were science-based but not specifically chemistry or physics and some, chiefly deduction items, were only incidentally connected with science, and could have been re-worded to involve almost any subject.

Results for the two forms were combined (producing, incidentally, a "Physical Science Critical Thinking Test" of 70 items, $\alpha = 0.80$) and items were grouped as chemistry (C), physics (P), science (S) and general (G). These four groups were treated as new subtests. Reliabilities and correlations are shown in Table 5.

TABLE 5

Rearrangement of items

<table>
<tr><th></th><th>C</th><th>P</th><th>P.E.B. Chemistry</th><th>P.E.B. Physics</th><th>S</th><th>G</th></tr>
<tr><td>C (23)</td><td>0.53</td><td>0.41</td><td>0.39</td><td>0.30</td><td rowspan="2">0.59</td><td></td></tr>
<tr><td>P (20)</td><td></td><td>0.52</td><td>0.39</td><td>0.36</td><td></td></tr>
<tr><td>S (12)</td><td>0.27</td><td>0.29</td><td>0.31</td><td>0.28</td><td>0.26</td><td>0.55</td></tr>
<tr><td>G (15)</td><td>0.34</td><td>0.50</td><td></td><td></td><td></td><td>0.52</td></tr>
</table>

The fact that items in a group were taken on different days renders this treatment suspect, but the results are at least worth comment.

For such short tests, internal consistency for C, P and G items is moderately good, but low for S items. C and P items correlate rather better than do the whole tests, but the coefficient seems to be rather lower than that for internal consistency. The data concerning correlations with other tests and examinations are given in Table 6.

It is likely that all scores are influenced by a common factor, presumably IQ. It would appear, therefore, that Form C measures a different ability from those involved in the other tests.

Conclusions

It would be premature to state firm conclusions based on the data so far obtained. The number of cases is not large, and data is being collected from a larger number in order to minimise uncertainty in the calculated correlations.

TABLE 6

Critical Thinking Tests — Correlations with other tests

	Chemistry Knowledge	TOUS	Chemistry-Impact on Society	Problem Solving	P.E.B. Chemistry	P.E.B. Physics
C	0.37	0.52	0.35	0.48	0.29	0.35
P					0.28	0.22

However, it would appear that workers attempting to devise tests of critical thinking based on the WGCTA approach must be aware of problems of validity arising from the choice of subject matter. Just what effect the subjects has on a person's ability to correctly assess statements cannot yet be described; the field is a practically untouched and probably difficult area of research.

References

ENNIS, R.H. A concept of critical thinking. *Harvard Educational Review*, 1962, **32**, 82-83.

FOLLMAN, J. Factor analysis of three critical thinking tests, one logical reasoning test and one English test. *Yearbook of the National Reading Conference*, 1969, **8**, 154-160.

RUST, V.I. Factor analysis of three tests of critical thinking. *Journal of Experimental Education*, 1960, **29**, 177-182.

WATSON, G. & GLASER, E.M. *Watson-Glaser critical thinking appraisal.* New York: Harcourt, Brace & World, Inc., 1964.