DOCUMENT RESUME

ED 189 100 TM 800 170

AUTHOR Forster, Fred: Ingebo, George
TITLE Linking Groups of Items. Portland Public Schools Occasional Papers in Measurement No. 19.
INSTITUTION Northwest Evaluation Association, Oreg.: Portland Public Schools, Oreg.
PUB DATE [78]
NOTE 23p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education: Equated Scores: Field Tests: *Item Banks: *Latent Trait Theory: Reliability: Scaling: *Test Construction: Test Items: Test Validity
IDENTIFIERS Item Calibration: *Rasch Model: *Test Linking

ABSTRACT

In Rasch terminology, two or more tests are linked when they are joined into a pool related to a single scale. When the same item calibrates at a low level on one test and a high level on a second test, it indicates that the item is low on the first test because the other items on that test are higher, and is high on the second test because the other items on that test are lower. The difference between the average of the calibrations of a subset of items on one test and the average for the same items on a second test is called the linking value. The accuracy of a linking value can be established by confirmation through the use of a third test, an algebraic process known as triangulation. The pattern of links between pairs of field tests is called a linking network. The purpose for the network is to insure that sufficient information will be available to calculate accurate linking values for each test to the final scale. Four types of linking networks can be used: the four square network, the three-by-three network, the double eight linking network, and the octagon linking network. (Author/BW)

PORTLAND PUBLIC SCHOOLS

OCCASIONAL PAPERS IN MEASUREMENT

OCCASIONAL PAPER NO. 19

LINKING GROUPS OF ITEMS

Fred Forster
George Ingebo
Portland Public Schools

631 N.E. Clackamas Street
Portland, Oregon 97208

LINKING GROUPS OF ITEMS

Fred Forster
George Ingebo
Portland Public Schools

General Purpose

Linking is the art of making big ones out of little ones, item pools that is. The Rasch calibration of a single test relates the items to an equal interval scale centered on that test. The power of the Rasch model is realized when two or more tests are joined into a pool related to a single scale—in Rasch terminology "linked" to each other. Since every student can't take every item it is necessary to find an alternative method for tying separate test administrations together. The following guidelines are intended to establish some principles for linking items and tests in a variety of practical situations.

Rationale

When the same item calibrates at a low level on one test and a high level on a second test, it provides information about the relative calibrations for the rest of the items on the two tests. The item is low on the first test because the other items on that test are higher, and is high on the second test because the other items on that test are lower. For example, as shown in Figure 1, item X calibrated as 190 on test A* and 210 on test B. Since the difference in these calibration values is due to the level of the other items relative to item X, the twenty point difference is an estimate of how much the calibrations on one test would have to be increased or decreased to bring them into line with the other test. As expected, adjusting test A to test B (A$\longrightarrow$B) involves the addition of 20 points to the calibration of each item on test A. In this case, an item with a calibration of 230 on test A would be assigned a calibration of 250 in relation to test B items, i.e., would be expected to calibrate at 250 if included on test B.

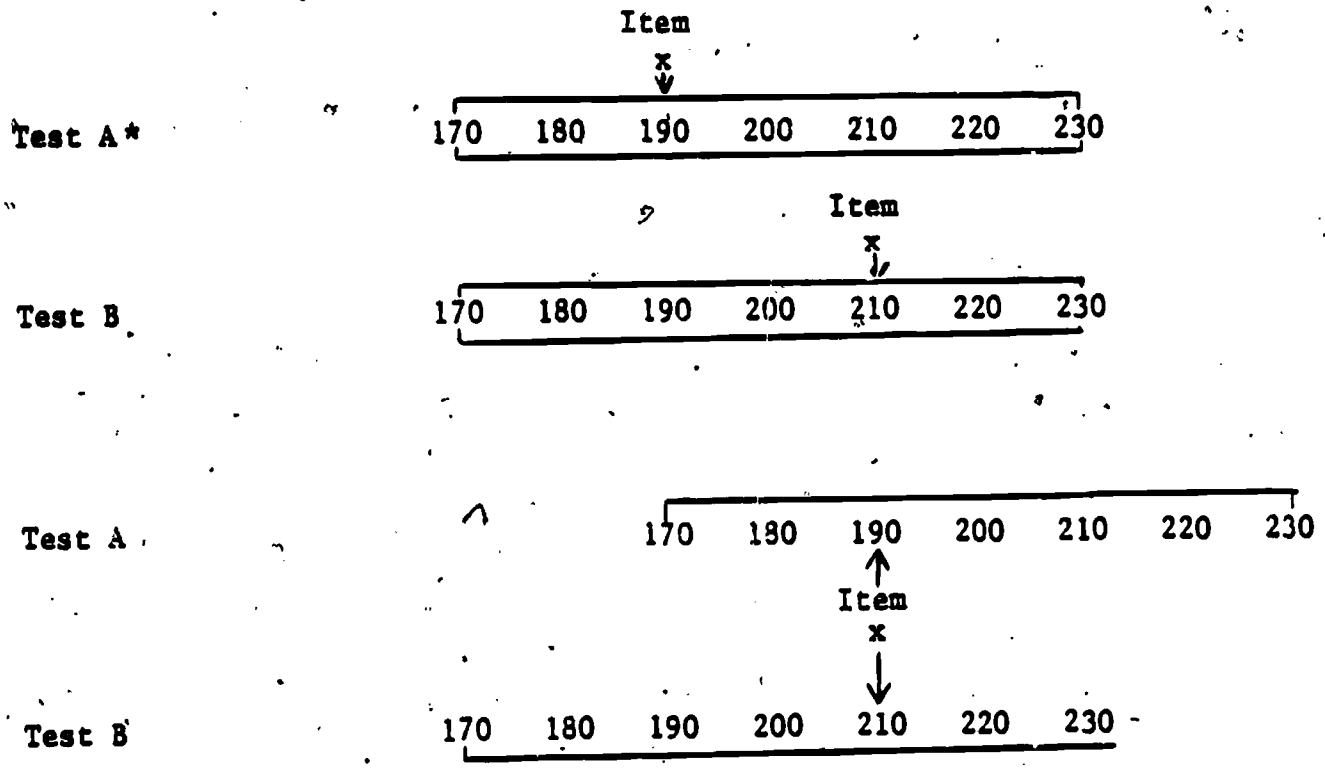
Obviously, one item is sufficient to link two tests if it is calibrated at the "right" value on both tests. In practice, the way to insure the stability of the linking value between two tests is to include a small subset of items on both tests. The number of items in this subset is one of several factors which are considered in planning efficient linking patterns.

---

*Calibrations are expressed in Rasch units (RITS).

Figure 1

The Same Item Calibrated on Two Different Tests



* As represented here, test A is at a higher level (more difficult) than test B.

Linking Values

A linking value is the difference between the average of the calibrations of a subset of items on one test and the average for the same items on a second test. This value can be either positive or negative depending on the levels of the tests being linked. As shown in Figure 2, the difference between the averages is +20 points if test A is adjusted to the scale established by test B (or -20 points if test B is adjusted to test A).

Linking values are more dependable as the number of items included in the subset or "link" is increased, but the size of the link does not guarantee accuracy. Since it is necessary to use a large number of tests to create an item pool, using a test length of forty items is about right to encourage teachers to volunteer testing time. Therefore, while an increase in the length of link increases the reliability of the linking value, it also significantly decreases the number of items being linked together.

Another factor affecting the size of a link is attrition. The Rasch program produces several indices of item quality which indicate the items that did not perform satisfactorily during field testing. Almost certainly link items will be lost in both the inital Rasch analysis and during the linking procedure. Therefore, more items need to be included in a link when field testing previously untried items, than when, for example, developing parallel forms of an established test.

Triangulation

The accuracy of a linking value is best established by confirmation through the use of a third test. As shown in Figure 3, adding the linking values for A⟶C and C⟶B should algebraically sum (approximately) to the linking value A⟶B. In this example the sum of A⟶C and C⟶B is called a confirming value for the direct value A⟶B.

The importance of triangulation is that it provides the capability of identifying and excluding counter-productive information from a linking value. In those cases where a close agreement is not obtained between the direct value and the confirming values for a link, some detective work is indicated and it may be necessary to reexamine the original Rasch scaling, the item analysis, or the linking analysis before the reason for the problem is uncovered.

Figure 2

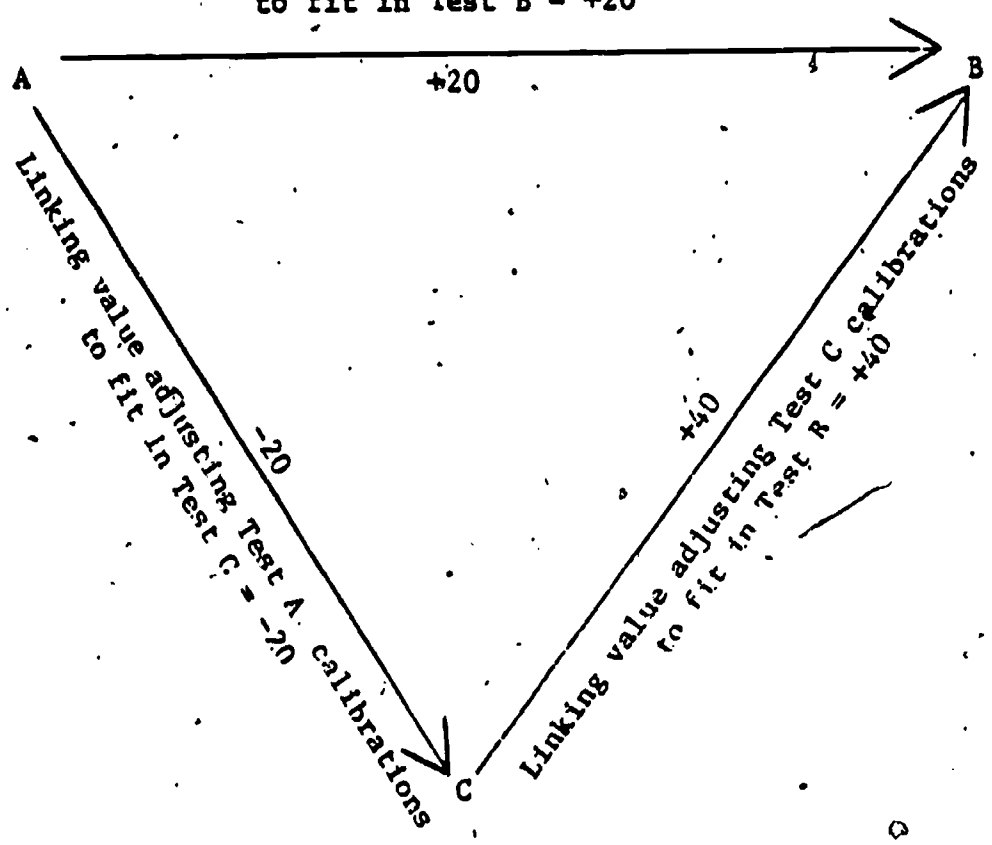
Calibrations for a Subset of Items on Two Different Tests

Item	Calibration* Test A	Calibration Test B
X	190	210
Y	187	206
Z	214	235
Average	197	217

* Calibrations are shown in Rasch units (RITS)

Figure 3

Example of the Triangulation of Linking Values



It is important to note that the philosophy behind triangulation implies that the discrepancies in linking values are caused by factors which should not be ignored. The linking network serves as a tool for identifying the discrepancies and the errors which caused them so that appropriate corrective action can be taken to remove the faulty data. This philosophy stands in contrast to some current linking procedures which average the discrepancies among links without any attempt to identify and eliminate possible sources of error.

Linking Procedures

The previous sections have described how linking establishes the item relationships which would exist if all items in a pool were calibrated in a single testing operation, and how confirming through triangulation validates the accuracy of the linking procedure. This section will describe a strategy for efficient linking. First, there will be an examination of the fundamental issues involved in constructing an item bank and field testing. Then there will be a consideration of the way in which several specific linking patterns provide the information needed to develop the scale for items being field tested.

Curriculum Considerations

Since curriculum integrity is essential to an item bank, it is necessary to carefully specify goals to which a bank relates and to precisely tie each item to these goals. In developing item banks in reading, mathematics and language usage, the Northwest Evaluation Association (NWEA) has used the comprehensive Tri-County Course Goal Collections for this purpose. (While NWEA efforts have proven effective in these subjects based on rules and conventions, item banking in other fields such as science and social studies where relationships are discovered rather than decreed remains to be proven.)

Curriculum balance is important in the design of field tests to make the exercise reasonable for teachers and students. A forty item field test which relates to the learning experiences of students can leave a positive reaction with a class, while a test which has too many items on the same goal or which is irrelevant to the goals of classroom instruction can be repetitious and boring. Curriculum decisions concerning the scope of content to be included in test and the number of items to represent each goal will affect the test format, test length, the number of tests needed, the composition of the links between tests, and the target population for

field testing. After the bank has been established and permanent tests are being built, decisions about the scope and sequence of content, the representation given different goals and the information needed from the test will help insure the success of the resulting testing program.

Administration Considerations

In addition to curriculum issues, it is important to consider a variety of administrative issues, especially as they affect the mental set of the teacher and students and, ultimately, the completeness and accuracy of the test information.

First, it is essential that the participation of principals and teachers in field testing be voluntary. We have found that one way to encourage volunteer participation is to return a report to the teather on a rapid turnaround schedule which is designed for classroom use. In addition to sparking interest, this policy increases the teacher's attention to the completeness and accuracy of information turned in for her class.

Second, the teacher volunteers need to be informed about the experimental nature of the tests, and that many of the test questions are new and untried. If they are encouraged to suggest improvements to the questions they don't like, an additional valuable source of information will be available for the item analysis.

Finally, it is important to use a test length for field testing which does not present difficulties for the teacher and, therefore, adversely affec her mental set and that of the students. After several years of experience, we have found that forty multiple choice items is about the maximum test length which does not present scheduling problems or introduce a speedness factor. While the maximum test length may differ from subject to subject, it is important to give the student sufficient time needed to complete the test if he knows the subject matter.

When these measures are followed, they lead to a better quality of test information and also contribute to the willingness of the teachers (as well as those with whom they discuss the program) to cooperate in future field testing.

Linking Networks

The pattern of item links (subsets) between pairs of field tests which is used to establish an item pool or expand an item bank is called a

linking network. The purpose for the network is to insure that sufficient information will be available to calculate accurate linking values for each test to the final scale thereby fixing the level of each item that has been field tested. Further consideration of this goal leads to a few general principles of network design.

First, it is usually not a good idea to use extremely large links. While it is true that linking values become more dependable as the size of the link is increased, a length of forty items is usually the maximum field test size which can be accommodated in normal administration. Together with the fact that the length of a link is only one factor affecting the accuracy of the linking value, this leads to the conclusion that using a link size beyond ten items usually unnecessarily reduces the item yield from field testing.

Second, it is necessary to assess the previous information available for your items. In addition to the number of items to be field tested and the existence of an established prior bank, two essential variables which affect a network are: (1) the number of independent calibrations needed for each item, and (2) the number of confirming values needed for each link. For example, in establishing an item pool with previously untried items each item should be calibrated at least three times and each link should be confirmed at least four times. On the other hand, if the items have been used and item analyzed before under good test administration conditions, then it might be reasonable to calibrate each item only twice and confirm each link twice.

Third, if the same link is used in all three tests involved in a triangulation, then the confirming value is of no use (see Appendix A).

Fourth, research on field testing has provided some rules of thumb for developing the linking network. Forster, et al (1978)* have shown that:

(1) it is not necessary to use random samples to obtain accurate item calibrations,

(2) 200 or more students are sufficient to obtain accurate item calibrations,

(3) item linking values conform to an equal interval scale, and

(4) item calibrations are not biased by levels of the other items included in the field test.

The following section describes the use of these guidelines in the design

---

*Forster, F., Ingebo, G.I. and Wolmut, P. The Rasch Model Monograph Series, The Northwest Evaluation Association, December 1978.

of linking networks related to situations actually encountered in building or expanding an item bank.

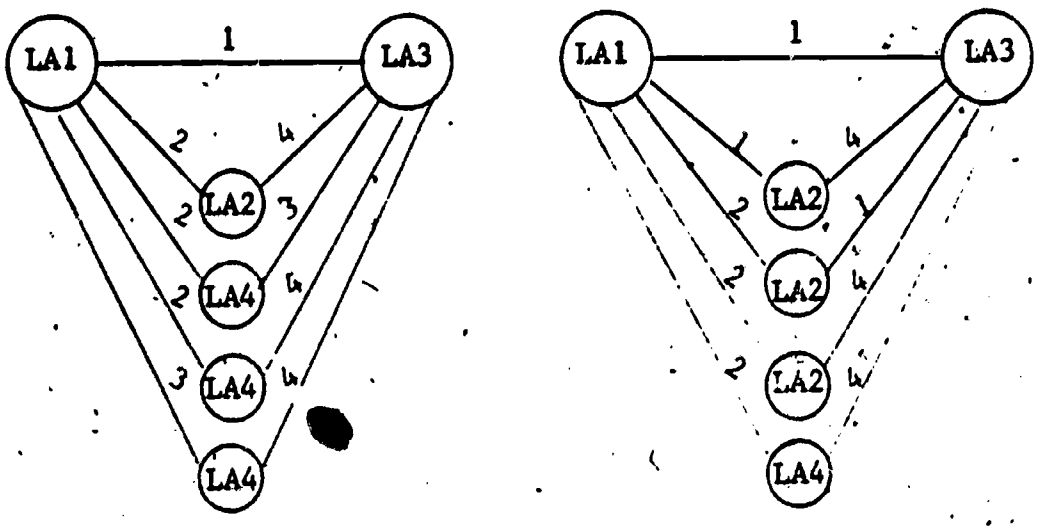
Examples of Linking Networks

The Foursquare Network

The first network discussed here was used to establish the NWEA language arts item bank from a large number of previously untried items. As shown in Figure 4, the Foursquare network is based on the principle of calibrating each item three times to produce two direct and four confirming values for each link, which makes it ideal for determing and excluding questionable data in building the bank. By numbering each block of ten items, the design for the field tests has been shown in Figure 5. The pattern includes 88 blocks of ten items each (880 items), arranged in 64 tests requiring approximately 13,000 students, giving a yield of 13.75 items per test.

As shown in Figure 6, this linking network derives its name from the fact that each group of four tests can be analyzed to give a pool of 80 items. Each group of four 80 item pools can then be analyzed to give a 240 item pool. The total of four 240 item pools can then be analyzed to give a single 880 item pool. Thus, there are three successive "phases" of the analysis leading to the final item pool.

In analyzing each of the links between two tests (say LA1 and LA3) the following triangulations can be made:



This configuration then provides two direct and four confirming values for each link. Once tests LA1 through LA4 are combined into a single pool,

Figure 4

Phase I of the Foursquare Linking Network

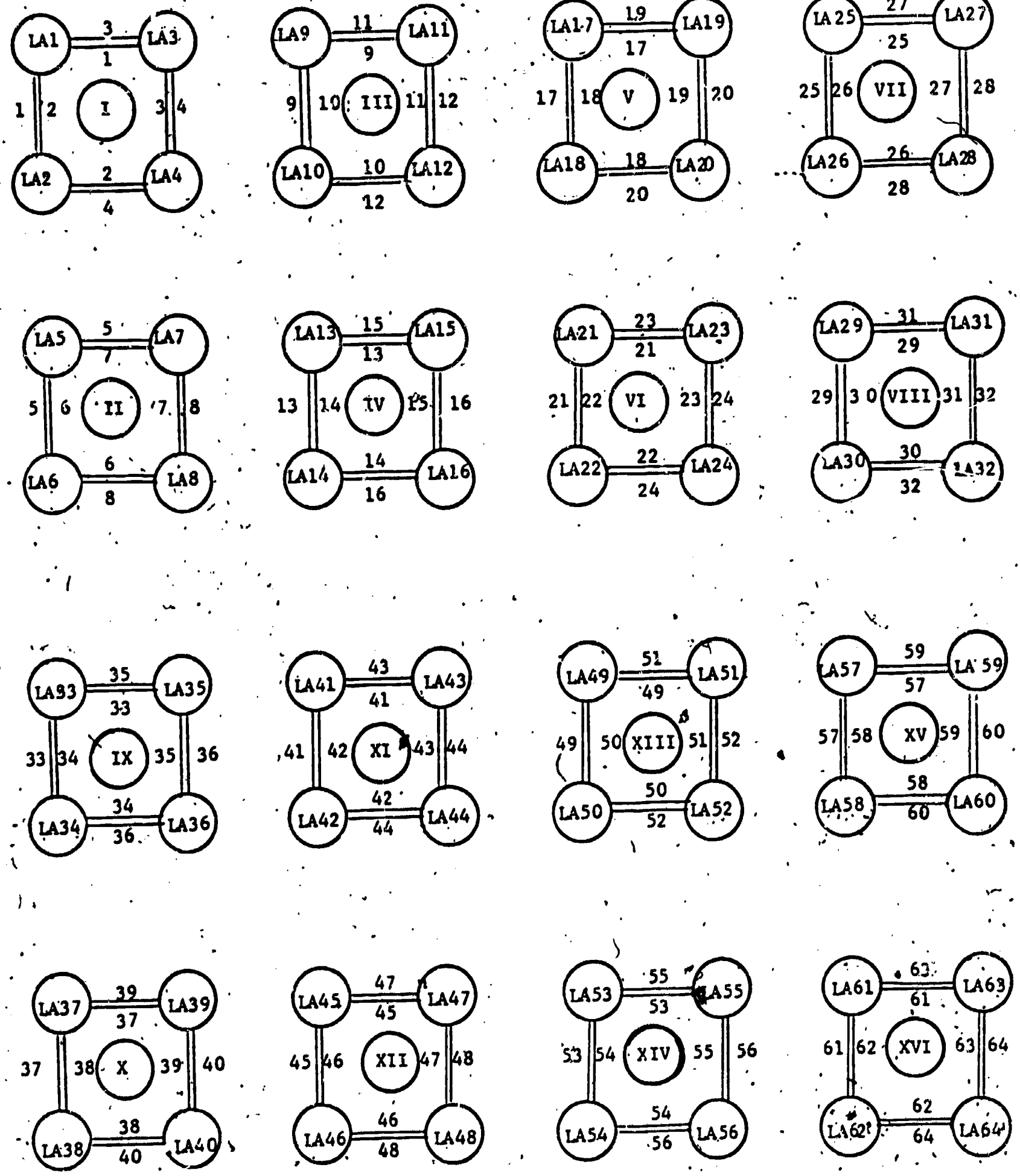


Figure 5

The Foursquare Linking Pattern

			Phase I								Phase II				
TEST	PHASE II	PHASE III	1	2	3	PHASE II	PHASE III	TEST	PHASE II	PHASE III	1	2	3	PHASE II	PHASE III
LA1	I	A	1	2	3	65	x	LA17	V	B	17	18	19	69	x
LA2	I	A	1	2	4	66	x	LA18	V	B	17	18	20	70	x
LA3	I	A	1	3	4	67	x	LA19	V	B	17	19	20	71	x
LA4	I	A	2	3	4	x	81	LA20	V	B	18	19	20	x	81
LA5	II	A	5	6	7	65	x	LA21	VI	B	21	22	23	69	x
LA6	II	A	5	6	8	66	x	LA22	VI	B	21	22	24	70	x
LA7	II	A	5	7	8	68	x	LA23	VI	B	21	23	24	72	x
LA8	II	A	6	7	8	x	82	LA24	VI	B	22	23	24	x	82
LA9	III	A	9	10	11	65	x	LA25	VII	B	25	26	27	69	x
LA10	III	A	9	10	12	67	x	LA26	VII	B	25	26	28	71	x
LA11	III	A	9	11	12	68	x	LA27	VII	B	25	27	28	72	x
LA12	III	A	10	11	12	x	83	LA28	VII	B	26	27	28	x	84
LA13	IV	A	13	14	15	66	x	LA29	VIII	B	29	30	31	70	x
LA14	IV	A	13	14	16	67	x	LA30	VIII	B	29	30	32	71	x
LA15	IV	A	13	15	16	68	x	LA31	VIII	B	29	31	32	72	x
LA16	IV	A	14	15	16	x	*85	LA32	VIII	B	30	31	32	x	*86

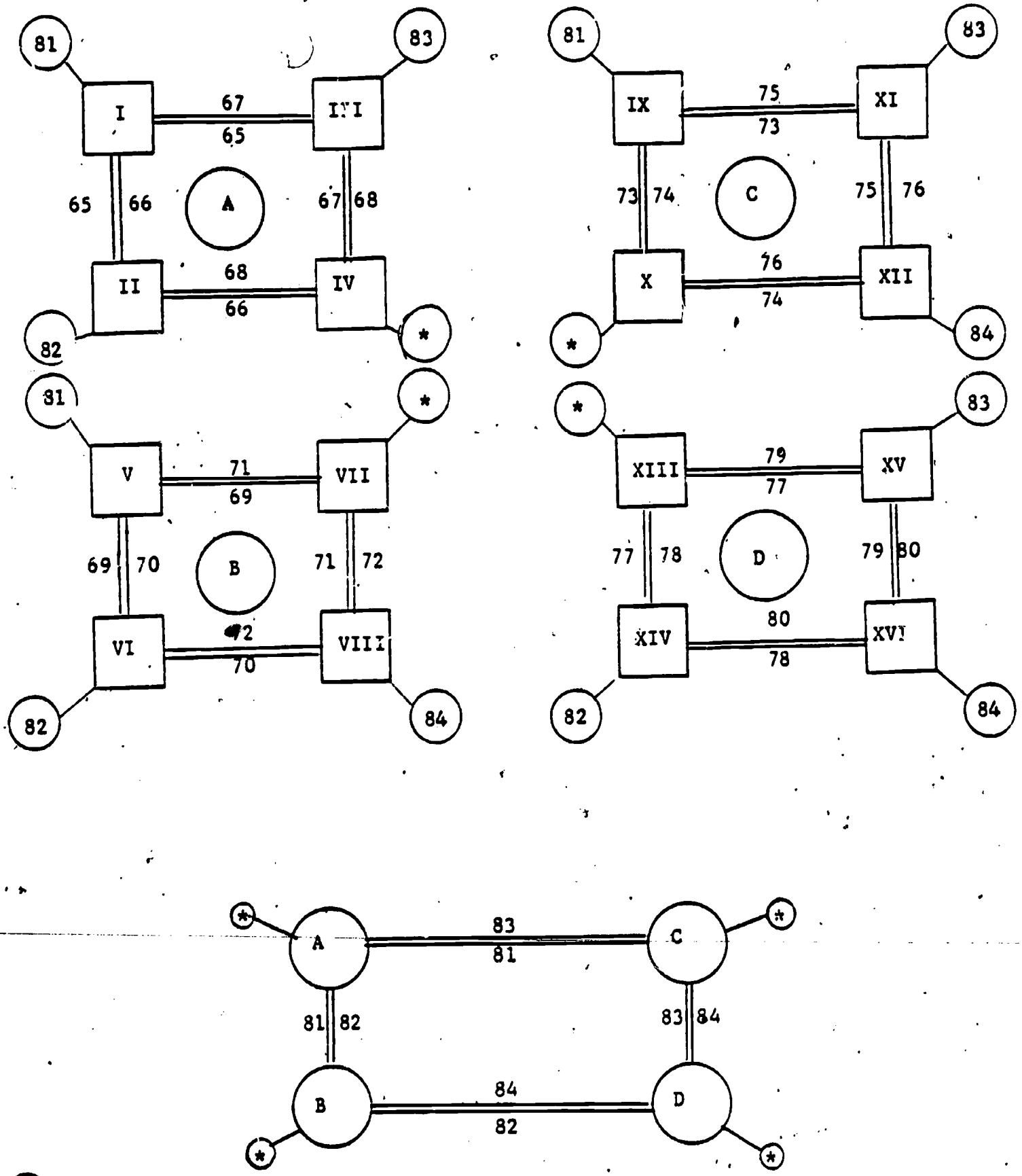
* Available for linking to higher levels.

			Phase I								Phase II				
TEST	PHASE II	PHASE III	1	2	3	PHASE II	PHASE III	TEST	PHASE II	PHASE III	1	2	3	PHASE II	PHASE III
LA33	IX	C	33	34	35	73	X	LA49	XIII	D	49	50	51	77	X
LA34	IX	C	23	34	36	74	X	LA50	XIII	D	49	50	52	78	X
LA35	IX	C	33	35	36	75	X	LA51	XIII	D	49	51	52	79	X
LA36	IX	C	34	35	36	X	81	LA52	XIII	D	50	51	52	X	82
LA37	X	C	37	38	39	73	X	LA53	XIV	D	53	54	55	77	X
LA38	X	C	37	38	40	74	X	LA54	XIV	D	53	54	56	78	X
LA39	X	C	37	39	40	76	X	LA55	XIV	D	53	55	56	80	X
LA40	X	C	38	39	40	X	83	LA56	XIV	D	54	55	56	X	83
LA41	XI	C	41	42	43	73	X	LA57	XV	D	57	58	59	77	X
LA42	XI	C	41	42	44	75	X	LA58	XV	D	57	58	60	78	X
LA43	XI	C	41	43	44	76	X	LA59	XV	D	57	59	60	80	X
LA44	XI	C	42	43	44	X	84	LA60	XV	D	58	59	60	X	84
LA45	XII	C	45	46	47	74	X	LA61	XVI	D	61	62	63	78	X
LA46	XII	C	45	46	48	75	X	LA62	XVI	D	61	62	64	79	X
LA47	XII	C	45	47	48	76	X	LA63	XVI	D	61	63	64	80	X
LA48	XII	C	46	47	48	X	*87	LA64	XVI	D	62	63	64	X	*83

*Available for linking to higher levels

Figure 6

Phase II of the Foursquare Linking Network



* indicates item blocks which may be used to link the network to an item bank or another network.

blocks 65, 66, and 67 link to the four adjacent pools (each formed by a group of four tests) through a comparable analysis.

The redundancy of direct links in this network strengthens the design and simplifies the identification of bad test information in any given test. Since organizing a field testing effort of this magnitude may leave little time available for readministration of a test shown to be weak, the redundancy provides a second direct value if one turns out to be unacceptable. If both direct values are unacceptable, there are still six more confirming values to salvage the link. When these failsafe features of the network are not needed, the two 10 item links between each pair of tests are combined to form six strong 20 item links among the four tests. Another feature of this network is that when sixteen or more tests are used it can accommodate a wide range of item levels. (In the case of the NWEA language arts bank, the 64 tests spanned grades three through eight.)

Theoretically there is no upper limit to the number of items which can be calibrated using this design. In addition, it can be modified to produce "unbalanced" networks so that a group of four tests can easily be linked to a group of sixteen or sixty-four.

The Three-by-Three Network

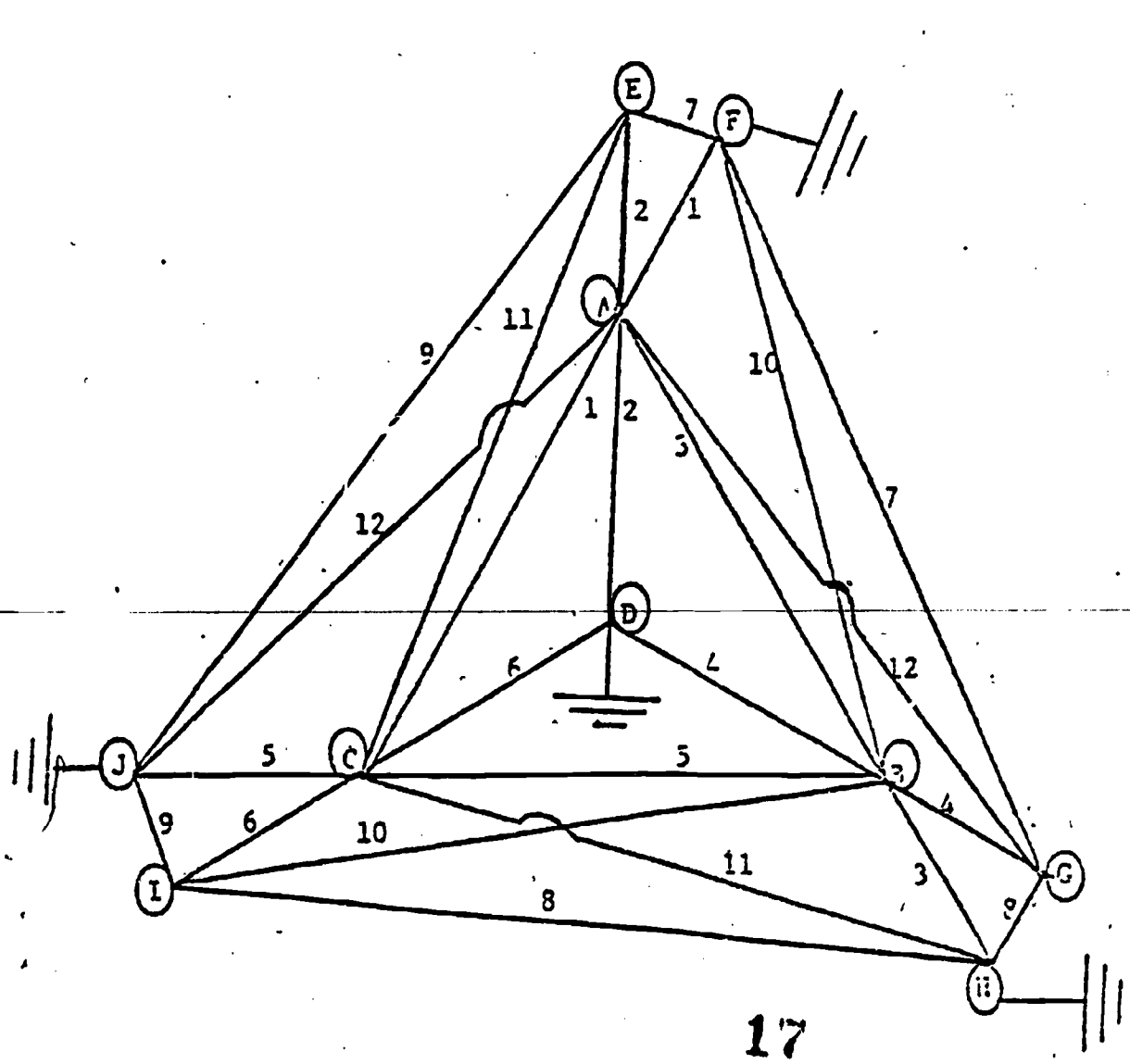
The Three-by-Three network is used when both new untried items and some established items are available for forming a pool or linking to an existing item bank. As shown in Figure 7, this network accommodates ten tests. Assuming that test A is the lowest level test and test J is the highest, care must be taken in selecting the items for link 12 since it is included in both of these tests. Therefore, one of the considerations in using this design is a determination of the appropriateness from a curriculum perspective of using the same items in the highest and lowest tests. (Normally, this constraint does not seriously limit a linking effort since the three bank links provide solid confirming values for this tenuous link.)

Using this design each new item is calibrated three times and each linking value is confirmed four or more times providing a yield of 120 new items or 12 items per test.

Figure 7

The Three-by-Three Linking Pattern and Linking Network

Test	10 Item Blocks			
A	1	2	3	12
B	3	4	5	10
C	5	6	11	11
D	2	4	6	Bank 1
E	2	7	9	11
F	1	7	10	Bank 2
G	4	7	8	12
H	3	8	11	Bank 3
I	6	8	9	10
J	5	9	12	Bank 4



17

||- indicates established or bank items.

The Double Eight Linking Network

The Double Eight network provides a flexible structure for accommodating mixes of proven and untried items. As shown in Figure 8, this network uses sixteen tests arranged in ascending level (from test A to test P) with four links to an existing bank. Each item is calibrated twice and each link has one direct and two confirming values. The sixteen tests link 360 items for a yield of 22.5 items per test.

One of the advantages of this design is its flexibility. In those situations where it is necessary to strengthen the pool by more tightly linking it to an existing bank, four additional bank links can be included. At the opposite extreme, if the timeline for field testing can be extended to accommodate readministering one or more forms, or nearly all the items will be included in a subsequent extensive testing program for recalibration, the links can be cut to eight items and a significantly higher yield will result. A third variation would be to repeat the design several times along parallel linking paths from a low level to a high level which would accommodate a wide range of curriculum scope and sequence and still provide adequate monitoring of the overall system.

The Octagon Linking Network

A shown in Figure 9, the Octagon network uses eight tests which are linked to an existing item bank and is appropriate when the field test length can be increased to 45 items. (The pattern uses five blocks of nine items for each test.) Each item is calibrated three times and there are one direct and four confirming values for each link. The network links 108 items (12 blocks of nine) to the bank for a yield of 13.5 items per test.

A useful variation on this design is to repeat the Octagon design four items as shown in Figure 10, and to use the "Bank" links to tie the four networks together. (This results in a structure quite similar to the Four-square pattern at the center of the four octagons.) The resulting network can be analyzed in two phases, first to create pools from each octagon of eight tests, and then to link the four pools with two direct and six confirming values for the links between pools.

Figure 8

The Double Eight Linking Network and Linking Pattern

A	1	2	3	4	Bank 1
B	1	9	10	11	12
C	5	9	13	14	Bank 1
D	4	5	6	7	8
E	2	10	13	17	18
F	3	6	14	15	16
G	12	17	23	24	Bank 2
H	11	18	25	26	Bank 2
I	7	16	21	22	Bank 3
J	8	15	19	20	Bank 3
K	24	26	30	31	32
L	20	22	27	28	29
M	23	25	30	33	35
N	29	32	33	34	Bank 4
O	19	21	27	34	36
P	33	34	35	36	Bank 4

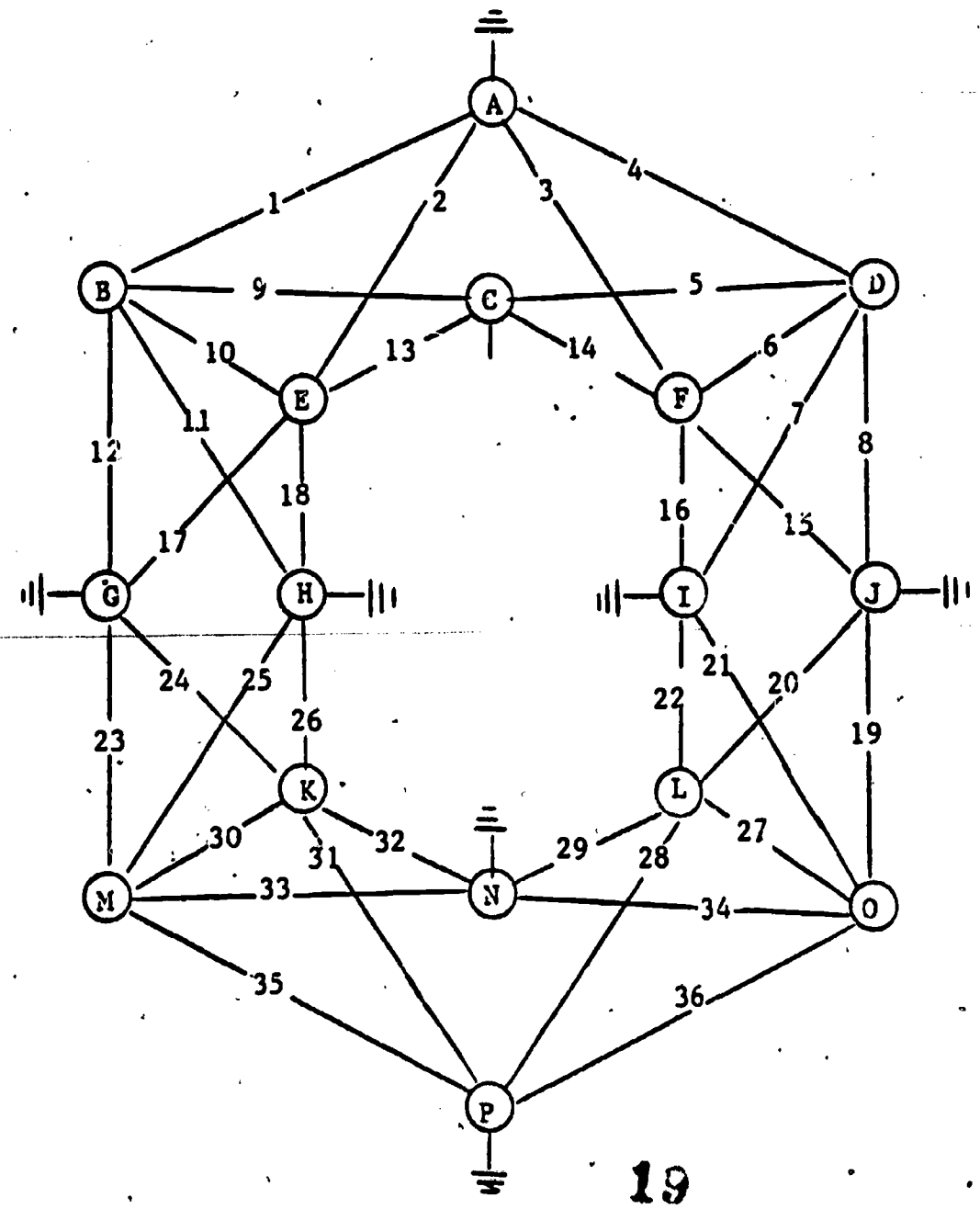


Figure 9

The Octagon Linking Network and Linking Pattern

A	1	8	9	6	Bank 1
B	2	4	5	10	Bank 2
C	1	2	7	10	12
D	3	5	6	9	11
E	2	3	8	11	Bank 3
F	4	6	7	12	Bank 4
G	1	3	4	9	12
H	5	7	8	10	11

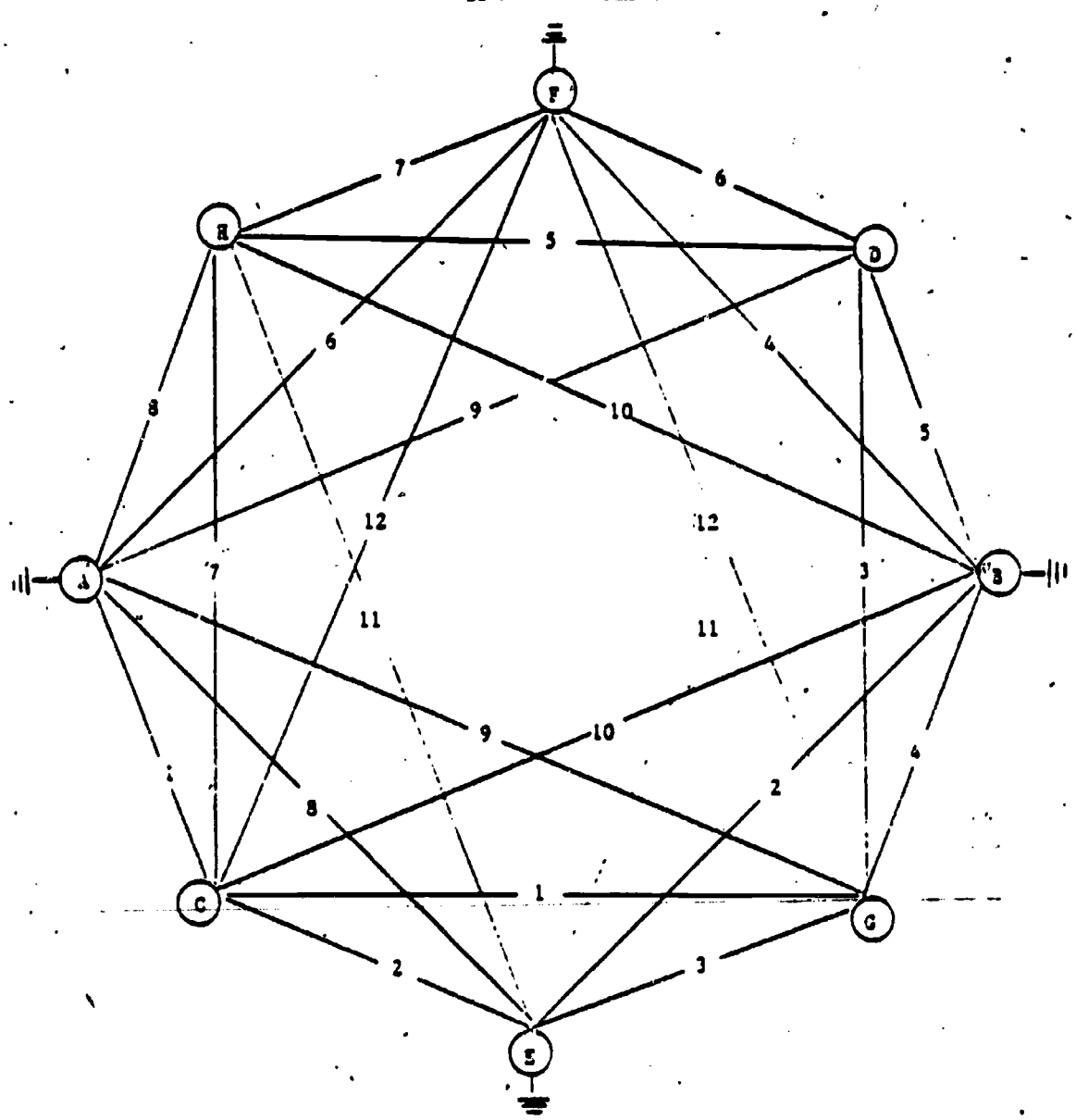
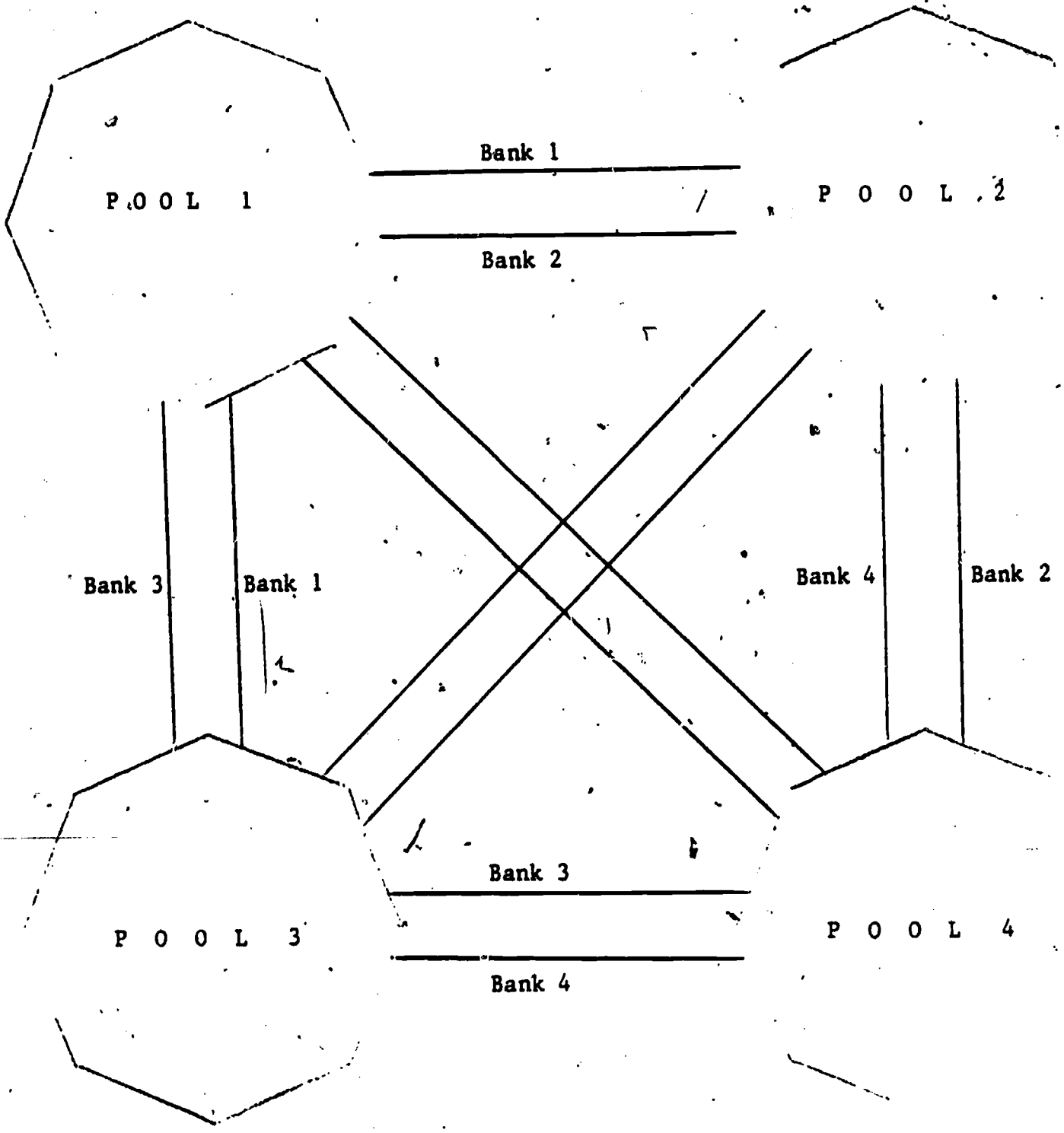


Figure 10

The Octagon Extended Network



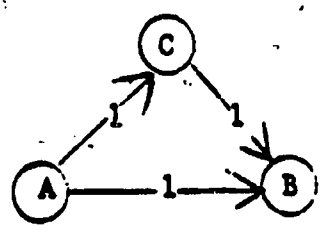
Summary

This paper has presented a rationale for linking tests to form item pools using Rasch calibration techniques and link triangulation. In addition, these principles have been demonstrated in the development of linking networks designed to meet different field test situations. The philosophy behind the development of these procedures is the desire to identify and eliminate faulty field test information. In actual practice we have encountered many instances in which an error in test administration which seriously biased item calibrations could only be identified by the comparison of the direct and confirming linking values in an appropriate network pattern. For this reason we believe the network approach to field test design is more practical and defensible than other currently popular procedures which average item calibrations without adequate regard for the influence of errors of validity and reliability.

Appendix A

Using the Same Link in All Three Tests of a Triangulation

(1) Assume Tests A, B, and C with linking block 1.



(2) Let 1_A be the average calibration for the items in block 1 on test A. Similarly for 1_B and 1_C.

(3) Then: $A \longrightarrow B = 1_B - 1_A$

$A \longrightarrow C = 1_C - 1_A$

$C \longrightarrow B = 1_B - 1_C$

(4) Then the confirming value is

$$(A \longrightarrow B) = (A \longrightarrow C) + (C \longrightarrow B) = (1_C - 1_A) + (1_B - 1_C) = 1_B - 1_C$$

(5) Since the confirming value must always be equal to the direct value $(A \longrightarrow B)$, the indirect link provides no confirming information.