DOCUMENT RESUME

ED 236 165 TM 830 619

AUTHOR Chevalaz, Gerard M.; Tatsuoka, Kikumi K.
TITLE A Comparative Analysis of Two Order Analytic Techniques: Assessing Item Hierarchies in Real and Simulated Data.
INSTITUTION Illinois Univ., Urbana. Computer-Based Education Research Lab.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CERL-CATM-RR-83-2-NTE
PUB DATE Apr 83
GRANT NIE-G-81-0002
NOTE 48p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Simulation; Instructional Design; *Item Analysis; Research Needs; Research Problems; Statistical Analysis; *Test Items
IDENTIFIERS *Hierarchical Analysis; *Item Hierarchies; Order Analysis; Ordering Theory

ABSTRACT

Two order theoretic techniques were presented and compared. Ordering theory of Krus and Bart (1974) and an extended Takeya's item relational structure analysis (IRS) by Tatsuoka and Tatsuoka (1981) were used to extract the hierarchical item structure from three datasets. Directed graphs were constructed and both methods were assessed as to how well they reproduced the theoretical structure of the data. It was discovered that the Krus and Bart (1974) procedure more adequately represented the complex interrelationships among test data than did the extended IRS method. Simulated data were found to present many problems and to be inappropriate for research in this area. Research in this area should include a large scale sampling distribution study to determine the distribution properties of simulated data. A more sophisticated method of generating binary responses which accounts for the distribution of theta needs to be developed. Also, a significance test and possibly a test of the differences between two item characteristic curves should be investigated. (Author/HFG)

Computer-based Education Research Laboratory CERL

A COMPARATIVE ANALYSIS OF TWO ORDER ANALYTIC TECHNIQUES: ASSESSING ITEM HIERARCHIES IN REAL AND SIMULATED DATA

GERARD M. CHEVALAZ

KIKUMI K. TATSUOKA

This research was partially supported by the National Institute of Education, under the grant No. NIE-G-81-0002. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

COMPUTERIZED ADAPTIVE TESTING AND MEASUREMENT

RESEARCH REPORT 83-2-NIE APRIL 1983

University of Illinois at Urbana-Champaign

Copies of this report may be

requested from:

Kikumi K. Tatsuoka
252 ERL
103 S. Mathews
University of Illinois
Urbana, IL 61801

Acknowledgement

Several of the analyses presented in this report were performed on the PLATO® system. The PLATO® system is a development of the University of Illinois, and PLATO® is a service mark of Control Data Corporation.

The authors would like to extend sincerest thanks to Dr. Maurice Tatsuoka for his assistance and kind cooperation in this sudy. Special thanks go to Bob Baillie for his helpful insights, Roy Lipshutz for his artwork, and Louise Brodie for aid in the preparation of this paper.

Abstract

Two order theoretic techniques were presented and compared. Ordering theory of Krus and Bart (1974) and an extended Takeya's item relational structure analysis (IRS) by Tatsuoka and Tatsuoka (1981) were used to extract the hierarchical item structure from three datasets. Directed graphs were constructed and both methods were assessed as to how well they reproduced the theoretical structure of the data. It was discovered that the Krus and Bart (1974) procedure more adequately represented the complex interrelationships among test data than did the extended IRS method. Simulated data was found to present many problems and to be inappropriate for research in this area.

Introduction

In order to correctly sequence blocks of instruction it is necessary to discover the underlying relationships between components of the instructional unit. Often it is important to uncover the hierarchical relationships of procedural tasks and to sequence instruction to facilitate learning. Tests can be used to discover this relationship. By assessing the relationships of test items, which reflect components of the instructional unit, educators can design and modify curricula. We can also check the extent to which we have succeeded in constructing problems that require a hierarchy of skills to be solved.

Methods for analyzing the relationships among items have existed for years. These include scalogram analysis (Guttman, 1950; Shevell, 1975) and Loevinger's (1947) analysis of item homogeneity. More recently however, methodologies have been developed to extract the best fitting hierarchy from test data.

The purpose of this study is to compare and assess two of these procedures, order analysis (Krus & Bart, 1974; Airasian & Bart, 1973) and item relation structure analysis (IRS) (Takeya, 1981). Both methods will be used to reconstruct a theoretical relationship among fraction addition test items.

Drawing from a combination of psychological measurement theory, formal logic theory, information theory, and graph theory concepts, order analysis and IRS present a general method of ordering two or more items. Both theories of discovering the hierarchical relationships among items can be divided into two components; 1) defining the order relation, and 2) extracting the item hierarchies.

Ordering theory has been developed to study hierarchical test structure. The hierarchical structure of a test is defined by a network of prerequisite relations among binary items (Bart, 1978). Binary data matrices are analyzed with respect to this relationship. The converse of the prerequisite relation is the dominance relation. If item i is a prerequisite to item j then item j dominates item i. The prerequisite or dominance relationhip is of primary interest in ordering theory. Briefly, a student is said to dominate an item if he/she passes that item, if he/she fails however, he/she is dominated by it. In the same manner, item i is a prerequisite to item j if for that student he/she answers item i correctly and item j incorrectly. In general, item i is said to be a prerequisite to item j if the percentage of students who answer item i correctly and item j incorrectly is greater than some constant.

Ordering analysis (Airasian & Bart, 1973; Bart & Krus, 1973) is a deterministic measurement model which expands scalogram techniques to assess nonlinear task networks. This model utilizes item response patterns to extract both linear and nonlinear prerequisite relations among tasks (Airasian, Madaus & Woods, 1975). Order analysis uses a set of primitive logic to isolate logical orders among variables in a hyperspace (Krus, 1978). The basis of an order relation, as defined by order analysis, is the characteristic of strong simple orders. Wise (1981) explains how strong simple orders have three properties:

asymmetry, connectedness, and transitivity. With regard to dominance, asymmetry implies that elements i and j cannot simultaneously dominate each other. Only one item can dominate the other. Connectedness, on the other hand, states that there must be a dominance relationship between two items i and j. The definition of transitivity allows implied item-item relationships. For elements i, j, and k within an order, if i dominates j, and j dominates k, then i dominates k.

In ordering theory all items must be dichotomously scored. If subject k answers item i correctly he/she is given a score of 1, while item i is scored 0 if subject k answers it incorrectly. Item i is then defined as a prerequisite to item j if the occurence of the response pattern (01) for items i and j is not found. Response patterns (00), (10), and (11), are referred to as confirmatory patterns and the pattern (01) is called a disconfirmatory response pattern (Bart & Krus, 1973; Airasian & Bart, 1975). Clearly the (00) and (11) response patterns do not provide any information as to whether item i is a prerequisite to item j.

Theoretically, there should be no inconsistencies of dominance. There should be no ij dominances for some students and ji dominances for others. However, even with unidimensional items such conflicting relations occur in practice due to measurement error. The manner in which item hierarchies are extracted and error in the data is delt with differs between the two order theoretic methods.

Bart and Krus (1973) originally attacked this problem in the following manner. For any set of items, a matrix which indicates the percentage of disconfirmatory response patterns for every pair of items can be produced. Every cell entry will be the percentage of times that a 0 for the i^{th} item and a 1 for the j^{th} item occurred. This table of percentages can be used to identify item pairs related by a prerequisite relationship. If the percentage of disconfirmatory patterns is less than a given tolerance level for any ij pair, then item i can be said to be a prerequisite to item j (Bart and Krus, 1973). The tolerence level sets the amount of disconfirmatory response patterns which will be allowed in defining the prerequisite relation. Finally, when the various prequisite relations have been defined, a hierarchy among the items can be constructed by applying the transitivity property. The hierarchical relationships among the items can be graphically represented by use of directed graphs.

More recently, however, McNemar's (1947) z statistic for comparing two correlated frequencies has been applied to analyze the prerequisite relations (Bart & Krus, 1973). As before, every element of a matrix is assigned a corresponding z_{ij} value where,

$$Z_{ij} = \frac{c-d}{(c+d)^{\frac{1}{2}}},$$

where c is the frequency of (10) patterns, and d is the frequency of (01) patterns. Again, a prerequisite relation is asserted if the percentage of disconfirmatory cases is less than the percentage of confirmatory cases. This translates into the condition that the corresponding z values exceed a predetermined alpha level. This removes chance prerequisite relationships due to measurement error.

The Japanese researcher Takeya, starting from the logic of Krus, Bart, and Airasian, has presented a different method of ordering called IRS. As with the Krus and Bart procedure, a binary data matrix is analyzed in terms of prerequisite relationships. Once again, the prerequisite relationship between items i and j is defined as success on item i is a prerequisite to success on item j. That is the response pattern (01) for items i and j respectively, does not occur. As before, the problem of the disconfirmatory pattern arises. Here Takeya's ordering approach departs from the Krus and Bart procedure.

Takeya (1980a, 1981) considers the statistical independence or dependence of scores obtained by two items. We denote a column vector of a data matrix X_{kj} by $\underset{\sim}{\theta}_j$ and its complement by $\bar{\underset{\sim}{\theta}}_j$, where

$$\bar{\theta}_{\sim j} = 1 - \theta_{\sim j} \quad .$$

If the proportion of correct and incorrect responses is expressed by

$$P(\theta_{\sim j}) = (1/N) \sum_{k=1}^{N} X_{kj} \quad ,$$

and

$$P(\bar{\theta}_{\sim j}) = 1 - P(\theta_j) \quad ,$$

then the proportion of subjects getting both items i and j correct is

$$P(\theta_{\sim i}, \theta_{\sim j}) = (1/N) \sum_{k=1}^{N} X_{ki} X_{kj} \quad .$$

The proportion of subjects getting item i incorrect and item j correct is

$$P(\bar{\theta}_{\sim i}, \theta_{\sim j}) = (1/N) \sum_{k=1}^{N} (1 - X_{ki}) X_{kj} \quad .$$

Takeya thus defines his coefficient of ordinality, r^*_{ij} , as:

$$r^*_{ij} = 1 - P(\bar{\theta}_i, \theta_j)/P(\bar{\theta}_i)P(\theta_j) \quad .$$

Table 1 reflects this relation.

Insert Table 1 about here

An IRS matrix is formed by calculating r^*_{ij} for all pairs of i and j. If r^*_{ij} is larger than a constant, the (ij)-cell is replaced by 1, otherwise 0.

However, unlike order analysis, Takeya's dominance relation does not satisfy the transitivity law. For example, if item i dominates item j, and item j dominates item k, item i does not dominate item k unless $r^*_{ik} > C$. By his definition of an order relation, implied item dominances are not allowed. Moreover, Takeya has not discussed an exact procedure for extracting the hierarchical relationships among items from the IRS matrix. So, Tatsuoka and Tatsuoka (1981) have proposed a procedure to extract directed graphs from the IRS matrix which uphold the transivity law. It is this modified IRS procedure which will be studied in this paper.

It should be noted that r^*_{ij} has a direct relationship to Loevinger's H_{ij}. Horst (1953) states that H_{ij} is an average ϕ/ϕ_{max}. Thus if we define a fourfold contingency table as

Table 1

Contingency Table of Items i and j

i \ j	1	0	total
1	$P(\underset{\sim}{\theta}_i, \underset{\sim}{\theta}_j)$	$P(\underset{\sim}{\theta}_i, \bar{\underset{\sim}{\theta}}_j)$	$P(\underset{\sim}{\theta}_i)$
0	$P(\bar{\underset{\sim}{\theta}}_i, \underset{\sim}{\theta}_j)$	$P(\bar{\underset{\sim}{\theta}}_i, \bar{\underset{\sim}{\theta}}_j)$	$P(\bar{\underset{\sim}{\theta}}_i)$
total	$P(\underset{\sim}{\theta}_j)$	$P(\bar{\underset{\sim}{\theta}}_j)$	1

$$\begin{array}{c|c|c|} & \multicolumn{2}{c}{i} \\ \hline j & a & b \\ \hline & c & d \\ \hline \end{array} \qquad \frac{c+d}{N} = P_j$$

$$\frac{b+d}{N} = P_i$$

with $b > c$

and $P_{i/j} = \dfrac{P(j+i)}{P_i} = \dfrac{d}{c+d}$.

Loevinger's H_{ij} can be shown to reduce to

$$\frac{ad - bc}{(a+c)(c+d)} = \frac{\phi}{\phi\ max} .$$

Moreover, by defining r^* in a similiar manner

$$\begin{array}{c|c|c|} & \multicolumn{2}{c}{i} \\ \hline j & a & b \\ \hline & c & d \\ \hline \end{array} \qquad \begin{array}{l} a+b = P(\bar{\theta}_{\sim i})N \\ c+d = P(\theta_{\sim i})N \end{array} \qquad \text{for } b < c .$$

$$a+c = P(\bar{\theta}_{\sim j})N \qquad b+c = P(\theta_{\sim j})N$$

Tatsuoka (1981) and Sato (1981) show that r^*_{ij} also reduces to

$$\frac{ad - bc}{(a+b)(b+d)} = \frac{\phi}{\phi\ max} .$$

Thus

$$r^*_{ij} = H_{ij} = \frac{\phi}{\phi\ max} .$$

Although Loevenger's work appeared first, H_{ij} was developed in another context and not applied to extracting hierarchical relationships among nonlinear task networks. For this reason the measure will be referred to as Takeya's coefficient of ordinality.

The purpose of this paper is to compare these two order theoretic methods and to assess which method more accurately extracts a theoretical hierarchical structure from binary data. More precisely, the order relation defined by ordering theory, and the method of extracting item hierarchies utilizing a given tolerance level of disconfirmatory responses (Bart & Krus, 1973) will be compared to the order relation defined by IRS and the chain extraction method developed by Tatsuoka and Tatsuoka (1981) which upholds transitivity. Graphs obtained by the Krus and Bart procedure and the extended IRS will be compared to the procedural network for fraction addition (Tatsuoka & Chevalaz, 1983) to see which best reproduces the theoretical hierarchy of fraction addition skills.

Method

Test and Subjects

Klein, et al. (1981) described the construction of a 48-item fraction-addition test for diagnosing erroneous rules resulting from misconceptions occurring at one or more levels of the procedural network. Klein and her associates constructed the test to consist of two parallel subtests. Each pair of items was constructed in terms of having identical procedural steps. The items reflect a variety of skills which are required to correctly add two fractions of varying types. Figure 1 is the procedural network for fraction addition as presented in Tatsuoka and Chevalaz (1983).

Insert Figure 1 about here

In an effort to assess and compare the Krus and Bart procedure and the modified IRS, the 48-item fraction test was administered to 148

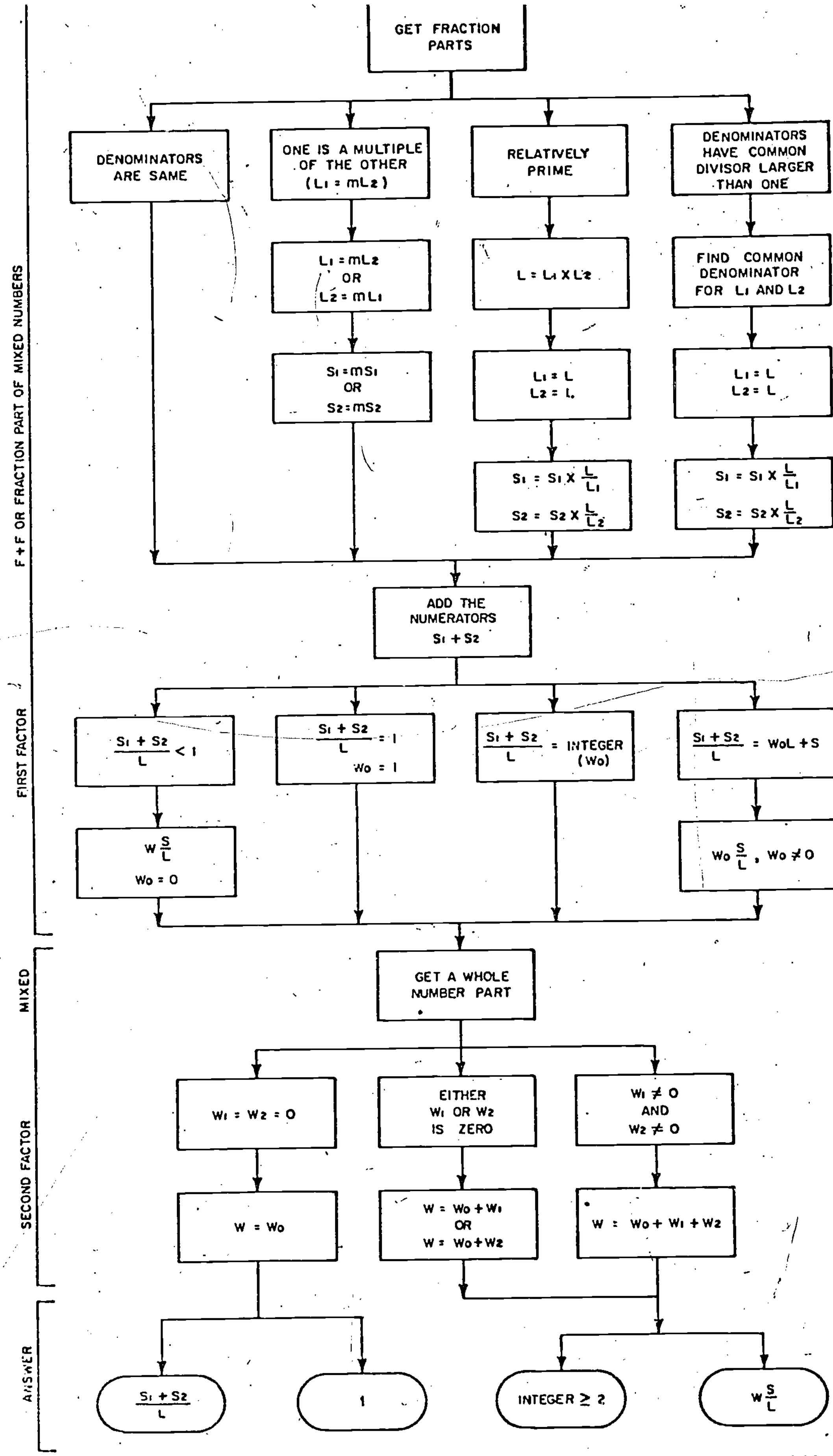


FIGURE 1: A Procedural Network for Fraction Addition

seventh and eighth grade students. After extensive logical error analysis (Klein, et al., 1981), and extraction of a undimensional subset of items by GETAB (Baillie, 1980), 36 items were retained for study. The estimated a's and b's of the two-parameter logistic model for the 36 items were calculated by GETAB (Baillie, 1982) along with the means and variances and are presented in Tables 2 and 3.

Insert Tables 2 & 3 about here

Datasets

Three different datasets, REAL, CLEAN, and SIM1, were employed in this study. Dataset REAL contains the binary responses for 148 students on 36 items. To avoid contamination by reducing task erors, the students' first nonreduced answer was chosen as his/her response. Each open-ended response was then converted into a decimal number and compared to a decimal number answer key. Items were given a value of 1 if the response and answer key matched and 0 otherwise. With this scoring procedure, choice of various common denominators or failure to reduce answers would not affect scoring.

Klein, et al., (1981) stated that there are two methods of solving fraction addition problems. The procedural network presented here, however, only reflects Method A of solving fraction addition problems. In this more commonly used method students add the whole number, denominator, and numerator parts separately. On the other hand, students who employ Method B first convert all mixed fractions to an improper fraction then add and reduce. Dataset CLEAN is a subset of REAL which consists of only those 119 students who used Method A when adding fractions.

Table 2
Estimated a and b Values for 36 Fraction Addition Items
From REAL (N = 148)

Item	a	b
1	.387	-.267
2	1.156	-.098
3	4.924	.368
4	2.756	.547
5	.523	-.968
6	1.656	-.419
7	2.972	.295
8	.738	-.490
9	1.561	-.734
10	2.562	.236
11	1.287	-.503
12	3.646	.406
13	1.166	-.402
14	8.637	.374
15	1.525	-.444
16	1.523	-.801
17	2.914	.398
18	1.996	-.302
19	1.100	-.391
20	1.336	-.386
21	4.819	.419
22	3.920	.554
23	1.493	-.195
24	1.591	-.431
25	4.287	.317
26	1.439	-.329
27	2.568	-.478
28	7.694	.428
29	2.206	-.348
30	3.579	.494
31	1.036	-.483
32	5.560	.399
33	1.221	-.523
34	1.500	-.637
35	4.579	.527
36	.927	-.532

Table 3

Means and Variances for 36 raction Addition Items (N = 148)

Item	μ	σ^2
1	.459	.520
2	486	.252
3	.419	.245
4	.318	.218
5	.574	.246
6	.581	.245
7	.426	.246
8	.534	.251
9	.635	.233
10	.439	.248
11	.581	.245
12	.392	.240
13	.554	.249
14	.432	.247
15	.581	.245
16	.642	.231
17	.384	.238
18	.568	.247
19	.457	.249
20	.561	.248
21	.399	.241
22	.324	.221
23	.527	.251
24	.581	.245
25	.432	.247
26	.554	.249
27	.608	.240
28	.412	.244
29	.581	.245
30	.351	.229
31	.561	.248
32	.412	.244
33	.581	.245
34	.615	.238
35	.345	.227
36	.561	.248

A simulated dataset, SIM1, was generated following a commonly used simulation procedure. First a pseudorandom number generator yielding a normally distributed set with mean 0 and variance 1 was used to simulate ability levels for 500 simulees. The probability that a given simulee would pass a specific item was given by

$$P_i(\theta) = \frac{1}{1+e^{-1.7a(\theta-b)}}$$

where a and b are the estimated a and b based on REAL and presented earlier (Lord, 1980). Next a random number between 0 and 1 was generated from a uniform distribution and compared to $P_i(\theta)$. If the probability of passing the item was greater than the random number, the simulated response was given a value of 1; conversely, if the probability of passing the item was less than the random number the simulated response was 0. In this manner 500 simulated response vectors of 36 items were generated.

To test the adequacy of SIM1 reproducing the qualities of REAL, GETAB was used to reestimate the item parameters. It was found, however, that the two-paramater logistic model would not converge for the simulated data. Furthermore, traditional item analysis showed that SIM1 differed greatly from REAL. To further look at the plausibility of using simulated data, four more simulated datasets, SIM2, SIM3, SIM4, and SIM5, were generated using different random number seeds. Again, the two-parameter logistic model would not converge for these datasets. The means, variances, and closest estimates of a and b for the 36 items and all five datasets are presented in Tables 4 and 5.

Insert Tables 4 & 5 about here

Table 4

Mean and Variance of 36 Items for Five Simulated Datasets

Item	Sim 1		Sim 2		Sim 3		Sim 4		Sim 5	
	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
1	.558	.247	.524	.250	.524	.250	.544	.249	.528	.250
2	.554	.248	.570	.246	.554	.248	.538	.249	.558	.247
3	.382	.237	.378	.236	.336	.224	.324	.219	.364	.232
4	.320	.218	.326	.220	.308	.214	.284	.204	.302	.211
5	.674	.220	.688	.215	.666	.223	.684	.217	.684	.217
6	.642	.230	.674	.220	.636	.232	.650	.228	.626	.235
7	.416	.243	.404	.241	.390	.238	.384	.237	.412	.243
8	.642	.230	.646	.229	.658	.225	.646	.229	.658	.225
9	.728	.198	.714	.205	.682	.217	.696	.212	.754	.186
10	.454	.248	.454	.248	.406	.242	.410	.242	.452	.248
11	.670	.222	.692	.214	.646	.229	.646	.229	.674	.220
12	.354	.229	.378	.236	.346	.227	.312	.215	.368	.233
13	.650	.228	.630	.234	.622	.236	.608	.239	.636	.232
14	.376	.235	.378	.236	.348	.227	.238	.221	.372	.234
15	.620	.236	.642	.230	.616	.237	.616	.237	.638	.231
16	.752	.187	.762	.182	.738	.194	.752	.187	.734	.196
17	.362	.231	.372	.234	.334	.226	.318	.217	.372	.234
18	.626	.235	.638	.231	.604	.240	.590	.242	.602	.240
19	.608	.239	.642	.230	.616	.237	.616	.237	.626	.235
20	.632	.233	.646	.229	.618	.237	.606	.239	.642	.230
21	.376	.235	.370	.234	.336	.224	.326	.220	.354	.229
22	.336	.224	.328	.221	.308	.214	.276	.200	.302	.211
23	.580	.244	.578	.244	.572	.245	.544	.248	.580	.244
24	.654	.227	.670	.222	.652	.227	.650	.228	.674	.220
25	.398	.240	.388	.238	.392	.239	.364	.232	.404	.241
26	.638	.231	.630	.234	.591	.241	.624	.235	.636	.232
27	.662	.224	.688	.215	.648	.229	.664	.224	.712	.205
28	.376	.235	.358	.230	.322	.219	.316	.217	.340	.225
29	.626	.235	.654	.227	.616	.237	.590	.242	.614	.237
30	.340	.225	.346	.227	.324	.219	.284	.204	.318	.217
31	.664	.224	.672	.221	.648	.229	.616	.237	.654	.227
32	.392	.239	.374	.235	.352	.229	.336	.224	.354	.229
33	.688	.215	.682	.217	.654	.227	.654	.227	.696	.212
34	.690	.214	.702	.210	.662	.224	.678	.219	.694	.213
35	.332	.222	.320	.218	.290	.206	.262	.194	.304	.212
36	.620	.236	.672	.221	.612	.238	.614	.237	.648	.229

Table 5
a and b Values of 36 Items for Five Simulated Datasets

Item	Sim 1		Sim 2		Sim 3		Sim 4		Sim 5	
	a	b	a	b	a	b	a	b	a	b
1	.237	-.228	.272	.087	.122	.350	.165	-.270	.069	.580
2	1.257	-.017	1.198	-.093	1.067	-.027	1.019	-.205	1.623	-.056
3	7.284	.504	8.035	.491	7.052	.535	6.073	.615	7.516	.404
4	3.172	.681	3.634	.635	3.583	.625	2.627	.759	3.511	.580
5	.288	-1.359	.336	-1.337	.410	-.925	.296	-1.514	.333	-1.294
6	1.470	-.313	1.618	-.424	1.496	-.268	1.548	-.384	1.590	-.259
7	2.857	.416	3.608	.430	4.384	.435	3.583	.469	3.304	.315
8	.551	-.585	.592	-.594	.533	-.674	.428	-.813	.637	-.616
9	1.408	-.653	1.492	-.590	1.723	-.401	1.437	-.565	1.889	-.651
10	2.687	.310	2.895	.297	3.504	.400	2.437	-.397	3.321	.221
11	1.086	-.481	1.110	-.576	1.271	-.330	1.159	-.419	1.281	-.454
12	3.921	.576	4.907	.494	5.820	.520	3.597	.656	6.306	.398
13	1.381	-.352	1.119	-.325	1.318	-.240	1.138	-.273	1.052	-.359
14	19.276	.523	15.287	.490	11.039	.512	2.004	.597	1.958	.383
15	1.629	-.220	1.388	-.332	1.675	-.189	1.097	-.309	1.510	-.303
16	1.875	-.689	1.336	-.822	1.851	-.587	1.556	-.771	2.150	-.557
17	3.672	.557	3.750	.512	5.079	.527	3.024	.648	3.945	.401
18	2.154	-.221	1.601	-.298	2.157	-.129	1.600	-.173	1.865	-.172
19	1.142	-.224	.930	-.415	1.192	-.235	.919	-.342	1.276	-.289
20	1.155	-.315	1.293	-.359	1.146	-.248	1.084	-.273	1.305	-.339
21	5.200	.517	7.621	.511	13.158	.529	5.035	.612	7.416	.423
22	4.117	.621	3.979	.626	6.280	.592	4.018	.748	4.836	.554
23	1.360	-.103	1.122	-.128	1.562	-.059	1.392	-.061	1.671	-.118
24	1.558	-.348	1.541	-.417	1.402	-.333	1.395	-.399	1.552	-.415
25	5.735	.465	4.433	.470	5.284	.431	4.126	.521	5.129	.329
26	1.419	-.304	1.354	-.293	1.634	-.129	1.505	-.296	1.648	-.285
27	2.190	-.339	2.296	-.433	2.564	-.248	1.840	-.415	2.831	-.454
28	13.009	.521	10.033	.540	2.004	.541	7.197	.631	10.408	.443
29	2.057	-.224	1.746	-.343	1.914	-.175	1.617	-.173	2.137	-.195
30	3.803	.614	6.250	.571	5.173	.567	3.901	.727	4.382	.521
31	1.035	-.468	.920	-.549	1.016	-.386	.865	-.354	.908	-.467
32	11.555	.487	6.235	.502	7.911	.506	6.094	.587	12.117	.416
33	1.044	-.569	1.083	-.541	1.220	-.367	.887	-.573	1.504	-.564
34	1.656	-.470	1.480	-.544	1.831	-.325	1.257	-.523	1.530	.485
35	5.833	.620	6.047	.636	11.861	.606	4.874	.776	5.687	.539
36	.979	-.298	.911	-.552	1.267	-.229	.973	-.322	.990	-.420

Clearly the choice of the random number generator seed or the "randomness" has a great effect on the results of the simulation. This counter-intuitive result warrants caution in the use of simulated data in quantitative research. However, SIM1, was arbitrarily chosen for inclusion in this study to determine if simulated data will reflect the hierarchical item patterns in real data.

Order Analytic Procedures

To determine the capability of the Krus and Bart (1974) and the Tatsuoka and Tatsuoka (1981) procedures for extracting item hierarchies, all three datasets were analyzed and compared to the procedural network. However, only the first subtest of 18 items will be included in the analysis. This will aid in the interpretation as the graphs will be less complex. The program ORDER2, written by Antonak, Bart, and Lele (1979), extracted prerequisite relationships by the Krus and Bart procedure, while the modified Takeya analysis was carried out by IRS (Baillie & Tatsuoka, 1981). A tolerance level of 5% was chosen for the Krus and Bart procedure based on recommendations in the literature (Airasian and Bart, 1975; Airasian, et al., 1975). Based on Takeya's guidelines (Takeya, 1980b) the cutoff for r^*_{ij} was set at .5.

Regression Analysis

Finally, a multiple regression analysis was performed to assess which, and to what extent, item characteristics influenced item difficulty, i.e., students' performance. Each item was dichotomously scored on 16 charactersitic variables, such as (1) fraction is of F+F type or (3) the denominators are the same. The variables were coded 1 if the item possessed that quality and 0 otherwise. Item difficulty, P_i, was selected as the criterion, and the 16 characteristic variables were selected as predictors. The 16 characteristic variables are presented in Table 6.

Insert Table 6 about here

Results

The outcome of the multiple regression analysis indicates that the linear combination of only five item characteristic variables account for 87% of the variance in item difficultly. Variables 3, 1, 10, 16, and 6, had a significant effect on students' performance. Table 7 presents these results.

Insert Table 7 about here

Only these five significant item characteristics will be represented in the directed graphs. By following the relationships reflected in the graphs between items with similiar and dissimiliar characteristics, we can determine the adequacy of the two procedures.

The directed graphs resulting from ORDER2 and IRS for dataset REAL are presented in Figures 2 and 3, respectively. Figures 4 and 5 are the resulting directed graphs for CLEAN.

Insert Figures 2, 3, 4 & 5 about here

Examination of the directed graphs leads to several observations. First, graphs obtained by ORDER2 for the two datasets are considerably more complex than those obtained by IRS. ORDER2 shows more intricate interrelationships among items on the test. Earlier it was shown that the two-parameter logistic model converged for dataset REAL satisfying

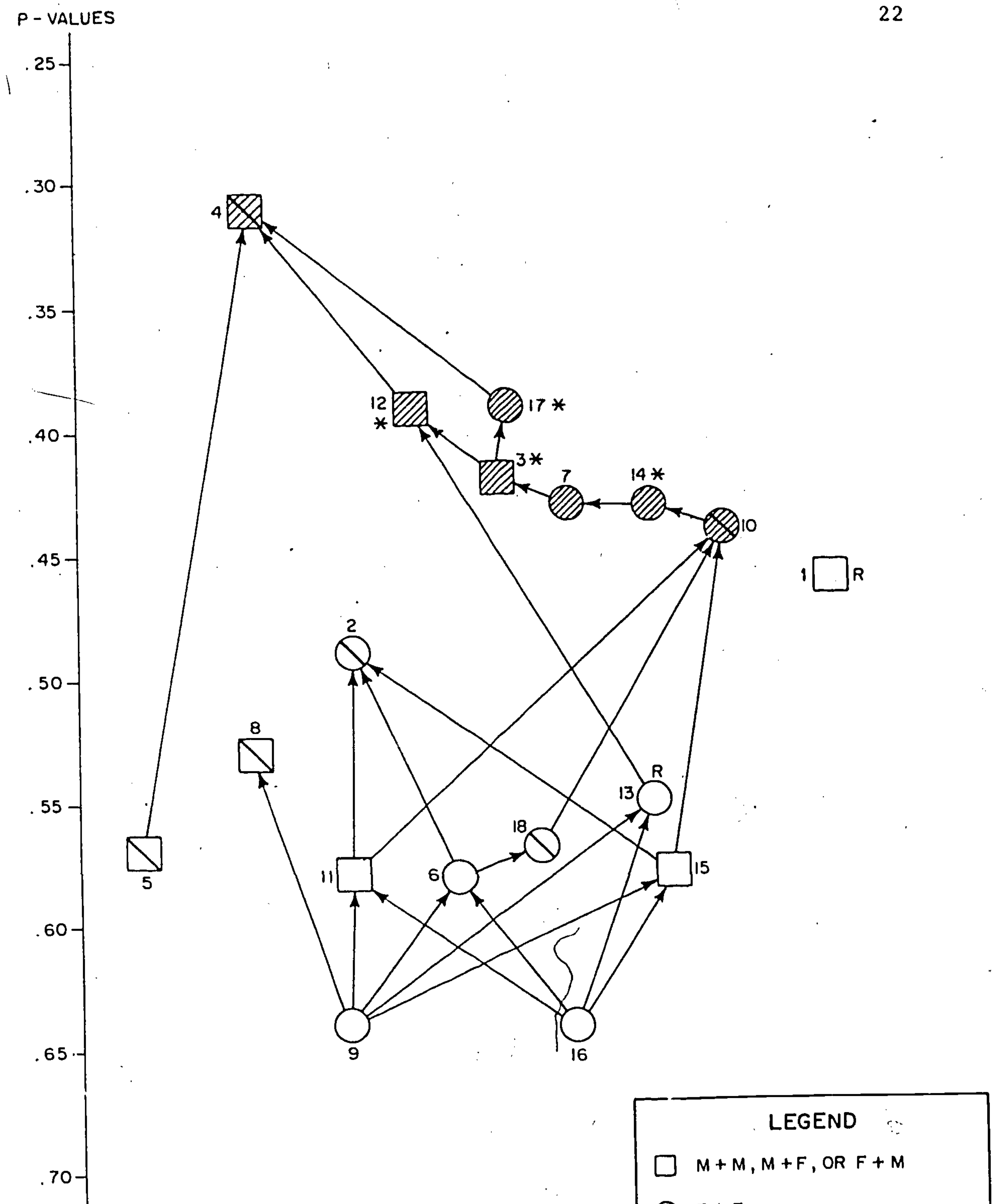
Table 6
Item Characteristic with Respect to Procedural Skills

Variable	Description
1	F+F $\frac{S_1}{L_1} + \frac{S_2}{L_2}$ or $\frac{S_1}{L} + \frac{S_2}{L}$
2	Mixed $w_1 \frac{S_1}{L_1} + w_2 \frac{S_2}{L_2}$, $w_1 \frac{S_1}{L_1} + \frac{S_2}{L_2}$, $\frac{S_1}{L_1} + w_2 \frac{S_2}{L_2}$
3	Denominators are same
4	One of the denominators is a multiple of the other
5	Two denominators are relative prime
6	Two denominators have a common divisor larger than one
7	$S_1 + S_2 < L$ (L is common denominator)
8	$S_1 + S_2 = L$
9	$S_1 + S_2$ is a multiple of L
10	$(S_1 + S_2)/L$ is a real number larger than 1
11	The answer needs reducing
12	the answer is a whole number
13	The answer is a mixed number
14	The fractions in a question can be reduced
15	One of the numerators is larger than L (common denominator)
16	Does second fraction need to be reduced?

Table 7

Regression of P_i on Five Item Characteristics with Respect to Procedural Skills

Multiple R	R^2	BETA Weights				
		Item Characteristic Number				
		3	1	10	16	6
.937	.878	.873	.243	-.315	-.335	-.114



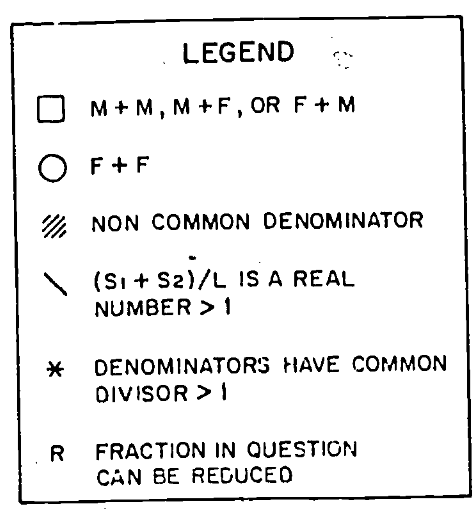


FIGURE 2: A Directed Graph of Real Data from Order 2

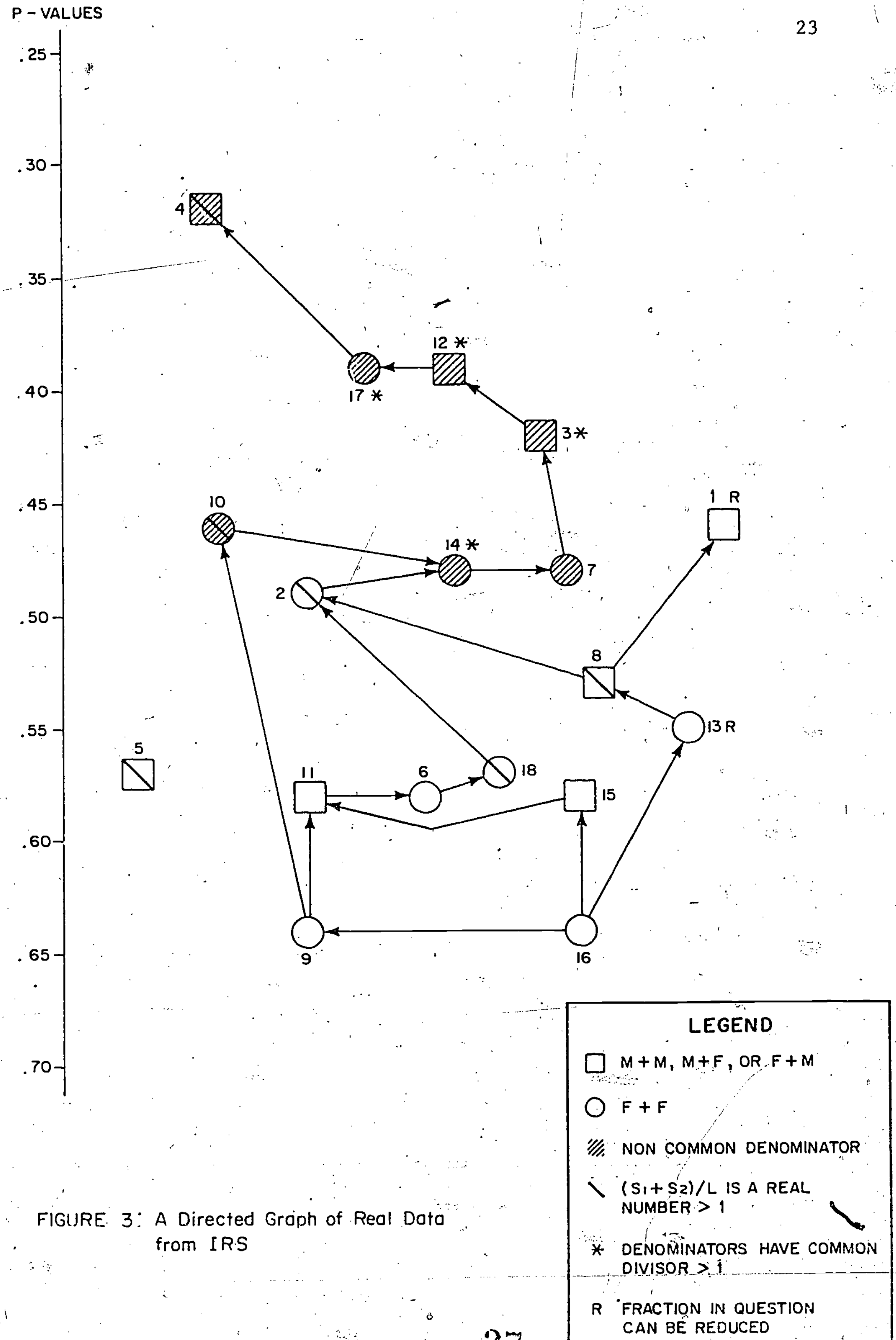


FIGURE 3: A Directed Graph of Real Data from IRS

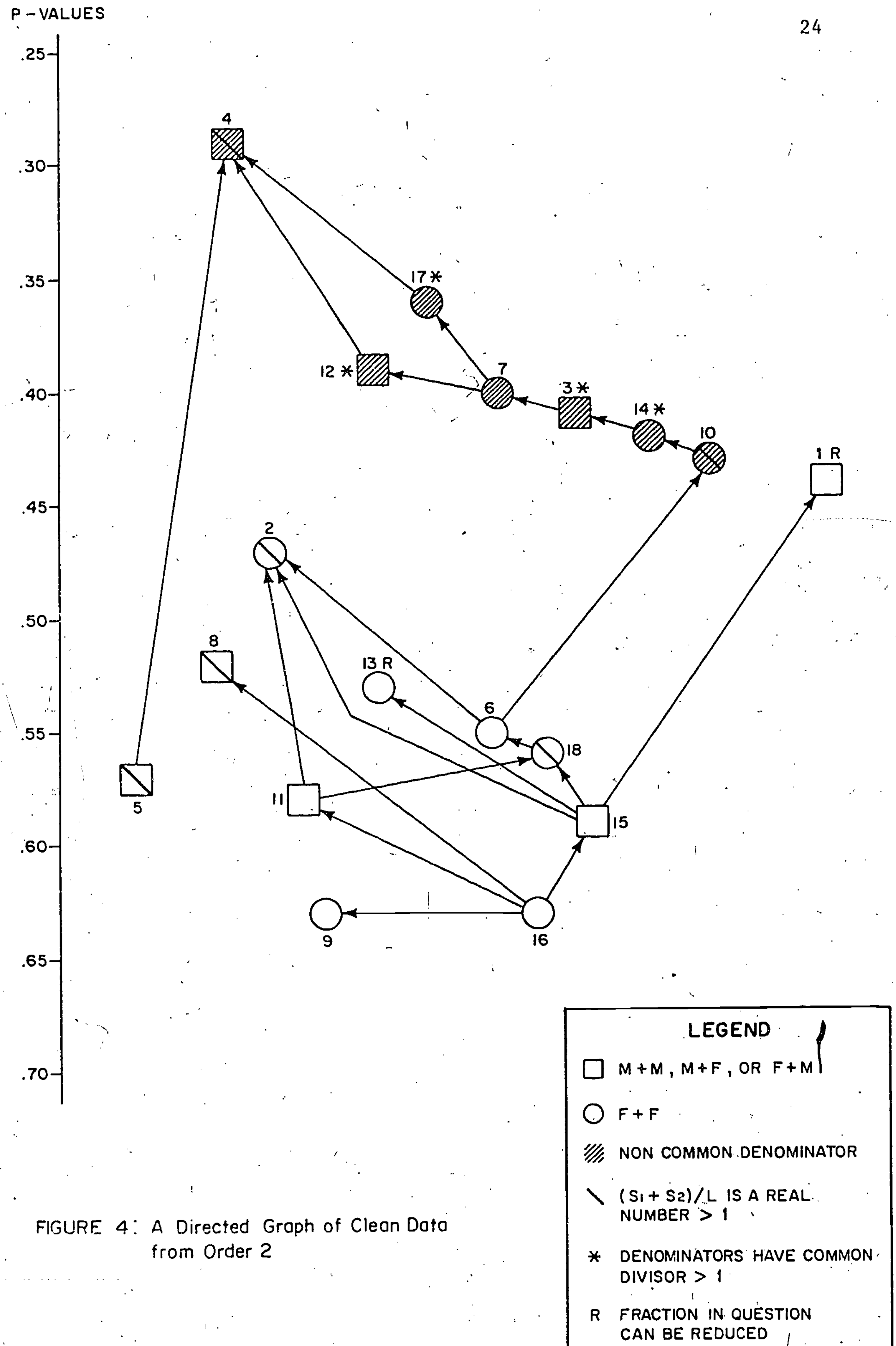


FIGURE 4: A Directed Graph of Clean Data from Order 2

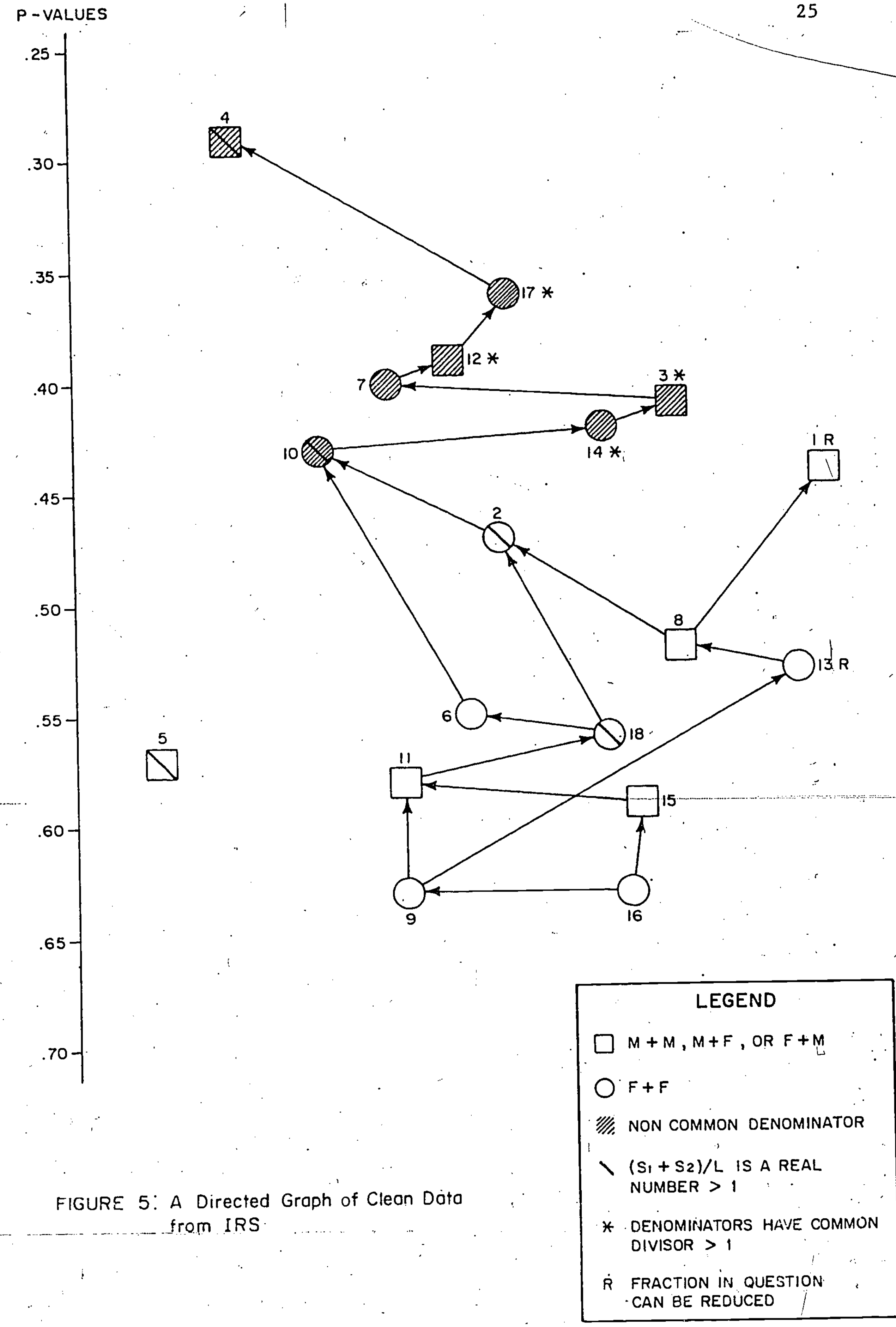


FIGURE 5: A Directed Graph of Clean Data from IRS

the assumption of unidimensionality. One would expect items of a unidimensional test to be highly interrelated. Comparason of Figures 2 and 3 reveal that by this criterion, ORDER2 more accurately expresses the data than IRS.

Both methods do a similiar job in separating out those items which have noncommon denominators from those which have common denominators. All graphs show that the procedural skills required to successfully complete common denominator problems are a prerequisite to the skills needed to correctly answer noncommon denominator problems. It is interesting to note that the common-noncommon attribute of an item appears to be the most influential aspect in determining students' performance. Noncommon denominator problems are not only more difficult; by both methods they appear to not be closely interrelated (connected in the directed graphs) with common demoninator problems. This, moreover, is a reiteration of the results of the multiple regression analysis and lends further credence to order analytic analysis.

The multiple regression analysis also demonstrated that the mixed fraction (M+M) vs. pure fraction (F+F) distinction was not significant in determining item difficulty. One would, a priori, have assumed that this would be a significant predictor. However, it must be kept in mind that the procedural network reflects only method A of solving fraction addition items. Since all parts of the fraction are added separately, conversion to an improper fraction is not required, and added procedural skills are not needed. In this sense M+M problems would not be much more difficult than F+F problems. Again, graphs from both ORDER2 and IRS reflect that fact. As discussed earlier, one would then assume if items of M+M and F+F type are similiar in nature then there would be many relationships or connections between and among these items. Again, ORDER2 appears to display this more fully.

The relationship between items of the type "$(S_1+S_2)/L$ is a real number greater than 1" is assessed differently by ORDER2 and IRS. IRS graphs for both REAL and CLEAN show a direct relationship between items of this type. Items 18, 8, 2, and 10, are all connected in a hierarchy. ORDER2 on the other hand, does not show this. Only in Figure 2, are two items of this similiar type related. Clearly, ORDER2 was not able to pick up this relationship among the items while IRS was.

In Figures 2, 3, and 5, items appear that are related to no other items by a prerequisite or dominance relation. IRS graphs for both REAL and CLEAN show that item 5 is not clearly dominated by any items nor does it dominate any other items. Furthermore, ORDER2 for REAL separated out item 1 from the other items. Intuitively this does not make sense; items 5 or 1 must be related to other items. Thus, these items must be of a nature (one that is not reflected in the graphs) such that students do not respond to them in any consistent manner. In this respect the performance on any other item is totally unrelated to performance on item 5 or 1. It is then a desirable quality of order analysis to separate out items of this nature.

In the Appendix is a copy of the 36 item test administered to the 148 students. Upon examining items 5 and 1, no salient characteristic appears that would make students respond in such a manner. IRT and classical test theory analysis do not flag these items. Single item groups and their relationship to the hierarchical structure of the test is an unanswered problem in order analysis.

Finally, the hierarchical relationhips between items in SIM1 are depicted in Figures 6 and 7. By first looking at Figures 2, 3, 4, and 5, and then at Figures 6 and 7, it quickly becomes apparent that the

Insert Figures 6 & 7 about here

simulated dataset, SIM1, did not reproduce the hierarchical structure among the 18 items. The graphical representation of this data further exemplifies the inflated higher mean values presented in Table 4. Neither the graph obtained by ORDER2 (Figure 6) nor that by IRS (Figure 7) are similiar to the graphs obtained by ORDER2 and IRS for REAL and CLEAN. All the interelationships among similiar items extracted by ORDER2 have been destroyed. IRS on the other hand, was able to extract a structure that is somewhat related to the structure of REAL.

It was hypothesized that the extreme a values reflected in Table 2 had a great effect on the ability of these two procedures to reproduce the observed data. To test this theory estimated a and b values from another dataset, REAL2, were calculated. REAL2 contains the binary responses of the 148 students for 36 items scored by a stricter scoring procedure. Each item was decomposed into its numerator part, denominator part, and whole number part. A response was scored 1 if each part of the response matched the three parts of the answer; otherwise it was scored 0. It should be noted that this scoring procedure necessitates that the student reduce his/her answer to form, else his/her response is marked incorrect. Since the procedural network

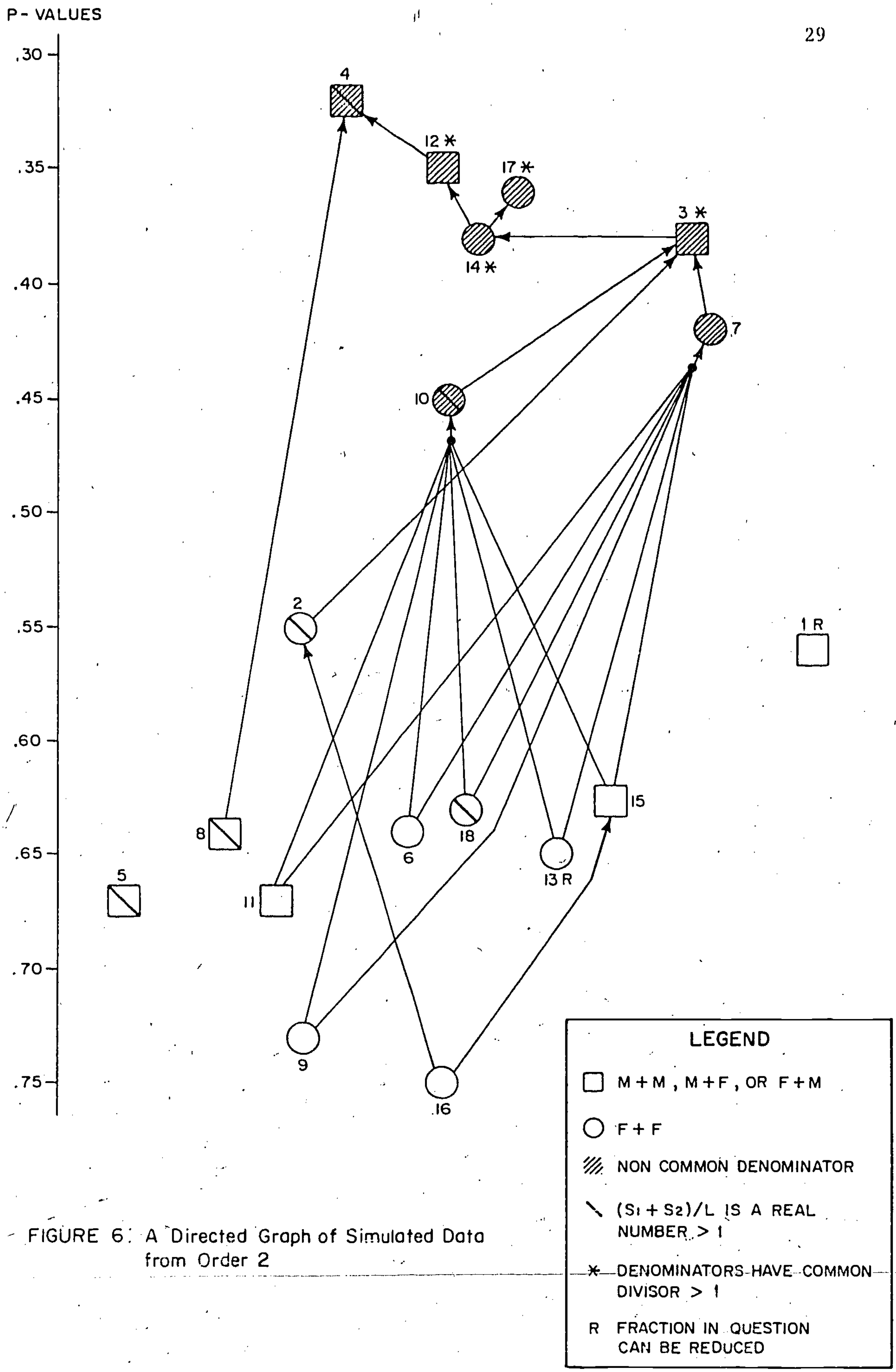


FIGURE 6. A Directed Graph of Simulated Data from Order 2

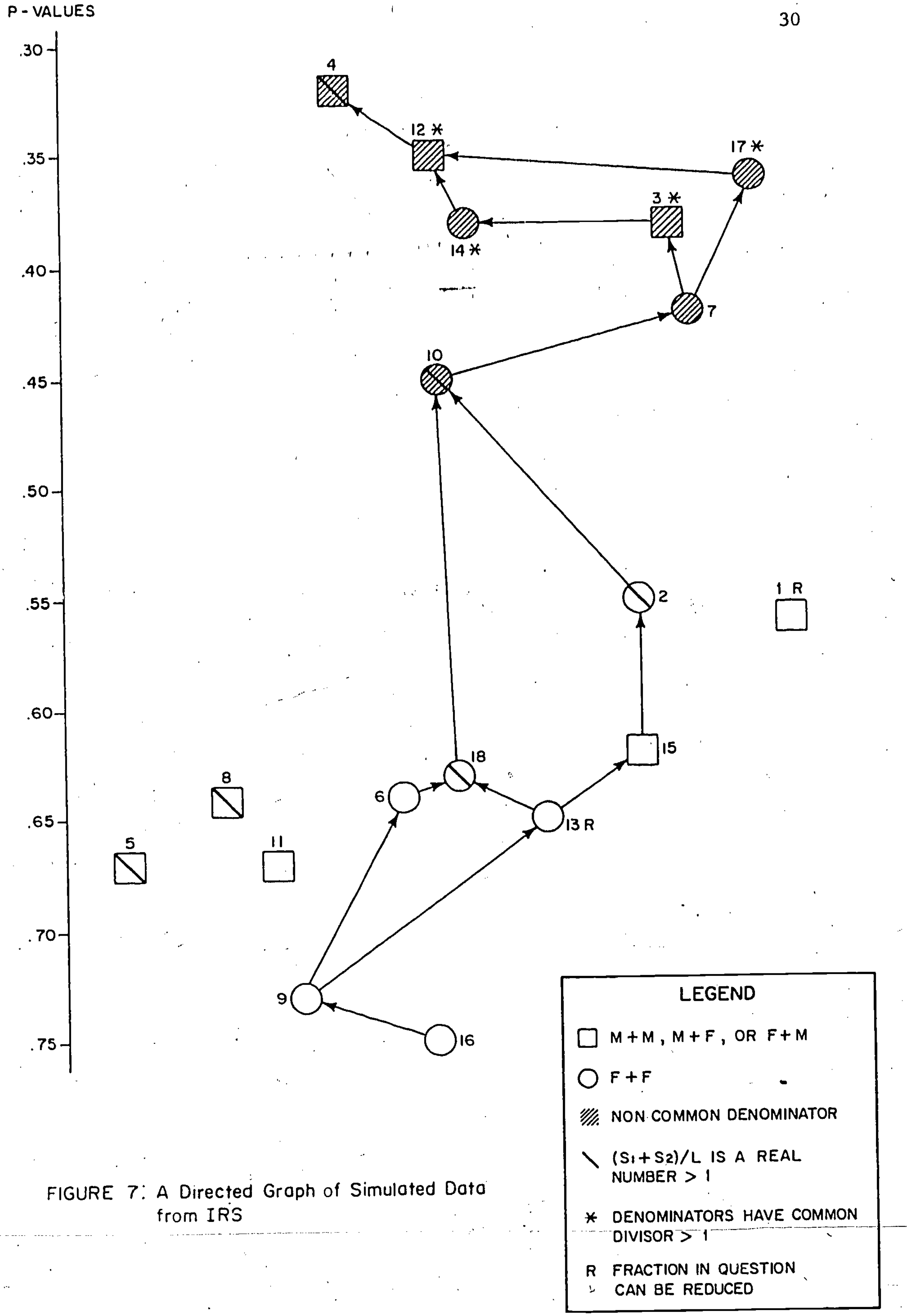


FIGURE 7: A Directed Graph of Simulated Data from IRS

does not account for reducing the resulting directed graphs of REAL2, REAL2 will not reflect the procedural network. Estimated a and b values for REAL2 were calculated by GETAB (Baillie, 1982) and are presented in Table 8. The means and variance of the 36 items are presented in Table 9.

Insert Tables 8 & 9 about here

These new estimated a and b values were then used to simulate 500 response vector. Once again, GETAB (Baillie, 1982) reestimated the item parameter of NSIM. The two-parameter logistic model converged for this data and the estimated a and b values are shown in Table 9.

Upon comparing Tables 8 and 10 it becomes quickly apparent that NSIM closely replicates the items characteristic of REAL2. However, if one compares Tables 9 and 11 again, great differences in the item means appear. These differences in item difficulties are reflected in great differences in the directed graphs. As before, the hierarchical structure of the original data is destroyed. Figures 8, 9, and and 11 display this result.

Insert Tables 10 & 11 about here

Insert Figures 8, 9, 10 & 11 about here

Clearly, this type of simulated data should be used with great caution in quantitative research which assess merits and shortcomings of various analyses. It was shown that not only can the choice of randon number generator seeds affect the data, but the quality of the original a and b values used in the simulation can have great affect on the results. Also, simulation data was shown not to maintain the hierarchical structure of the original data. The great differences in

Table 8
Estimated a and b Values for 36 Fraction Addition Items
From REAL2 (N = 148)

Item	a	b
1	.848	.754
2	1.594	.127
3	1.935	.153
4	2.028	.708
5	1.227	.279
6	1.823	.083
7	2.118	.037
8	.962	.364
9	.950	-1.158
10	1.882	.239
11	1.079	.045
12	1.700	.135
13	.884	.161
14	2.234	.009
15	1.563	-.275
16	2.042	-.930
17	1.977	.156
18	1.426	.173
19	1.498	.210
20	1.365	.198
21	3.368	.219
22	2.065	.708
23	1.382	.348
24	1.828	-.802
25	3.999	-.102
26	1.766	.201
27	.971	-1.240
28	2.879	.428
29	1.440	.205
30	1.610	.472
31	2.101	.686
32	2.490	.122
33	1.573	-.310
34	1.510	-.644
35	1.685	.422
36	1.225	.155

Table 9
Means and Variances for 36 Fraction Addition Items
from REAL2 (N = 148)

Item	μ	σ^2
1	.257	.192
2	.392	.240
3	.392	.240
4	.250	.189
5	.351	.229
6	.405	.243
7	.419	.245
8	.331	.223
9	.608	.240
10	.372	.235
11	.399	.241
12	.392	.240
13	.372	.235
14	.426	.246
15	.473	.251
16	.595	.243
17	.392	.240
18	.378	.237
19	.372	.235
20	.372	.235
21	.392	.240
22	.250	.189
23	.338	.225
24	.439	.248
25	.453	.249
26	.378	.237
27	.622	.237
28	.338	.225
29	.372	.235
30	.311	.216
31	.257	.192
32	.405	.243
33	.480	.251
34	.541	.250
35	.324	.221
36	.378	.237

Table 10
a and b Values of 36 Items Simulated Dataset NSIM (N = 500)

Item	a	b
1	.718	1.014
2	1.463	.086
3	2.128	.172
4	1.960	.878
5	1.041	.326
6	1.706	.134
7	2.042	.005
8	.711	.420
9	.553	-1.881
10	1.697	.226
11	.734	.044
12	1.773	.082
13	.611	.123
14	2.248	-.007
15	1.285	-.208
16	2.194	-.975
17	2.343	.199
18	1.223	.194
19	1.596	.183
20	1.042	.254
21	3.036	.212
22	1.785	.846
23	1.191	.262
24	1.560	-.188
25	3.406	-.141
26	1.742	.147
27	.877	-1.434
28	3.000	.392
29	1.030	.264
30	1.369	.586
31	1.394	.875
32	2.398	.138
33	1.349	-.472
34	1.337	-.701
35	1.727	.461
36	.963	.151

Table 11
Mean and Variance for 36 Fraction Addition Items from NSIM (N = 500)

Item	μ	σ^2
1	.308	.214
2	.484	.250
3	.454	.248
4	.256	.191
5	.426	.245
6	.468	.249
7	.508	.250
8	.422	.244
9	.796	.163
10	.440	.247
11	.500	.251
12	.484	.250
13	.486	.250
14	.512	.250
15	.568	.246
16	.790	.166
17	.444	.247
18	.456	.249
19	.454	.248
20	.444	.247
21	.438	.247
22	.268	1.970
23	.438	.247
24	.566	.246
25	.558	.246
26	.464	.249
27	.802	.159
28	.378	.236
29	.442	.247
30	.346	.227
31	.276	.200
32	.464	.249
33	.642	.230
34	.700	.210
35	.370	.234
36	.472	.250

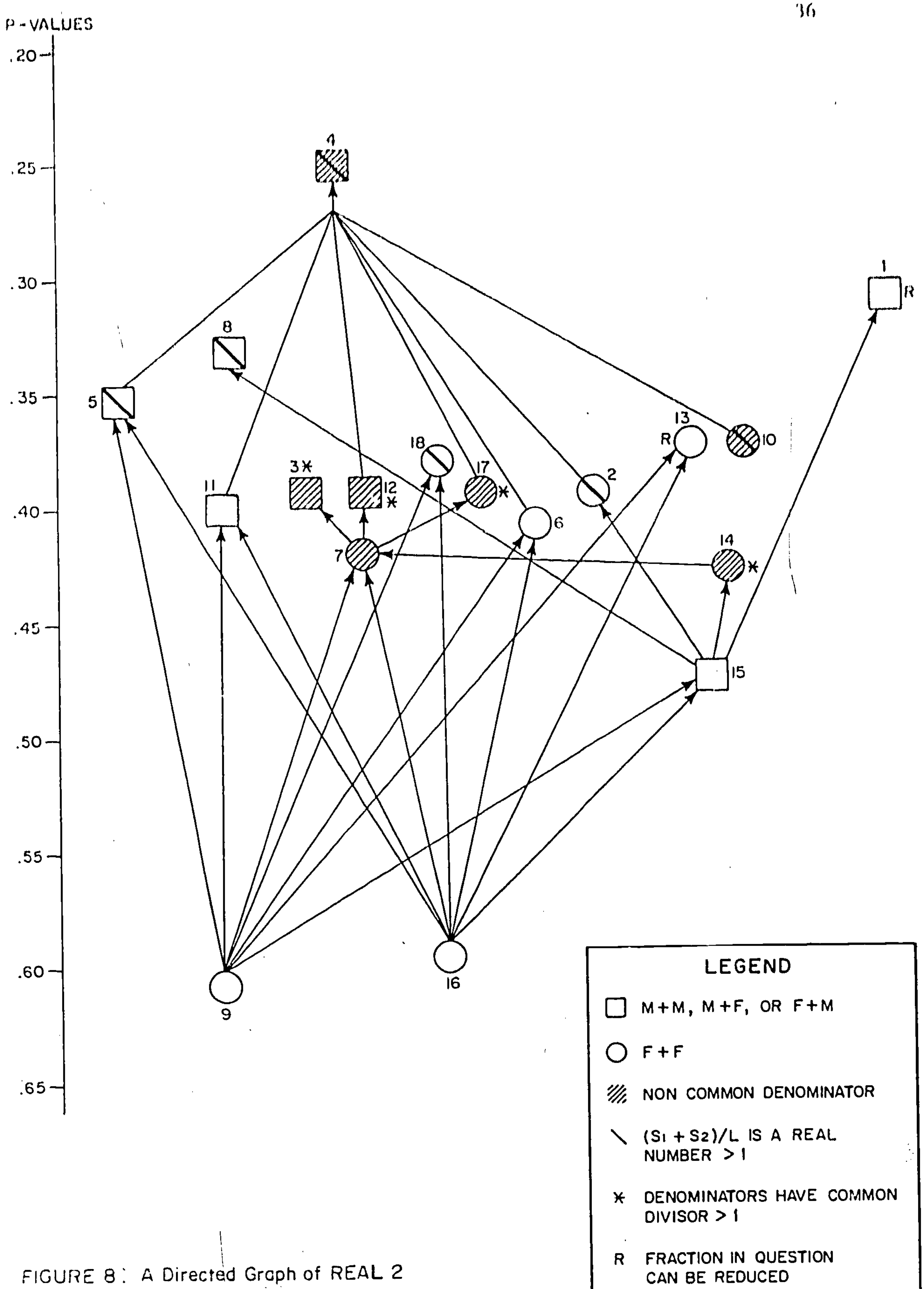


FIGURE 8: A Directed Graph of REAL 2 from ORDER 2

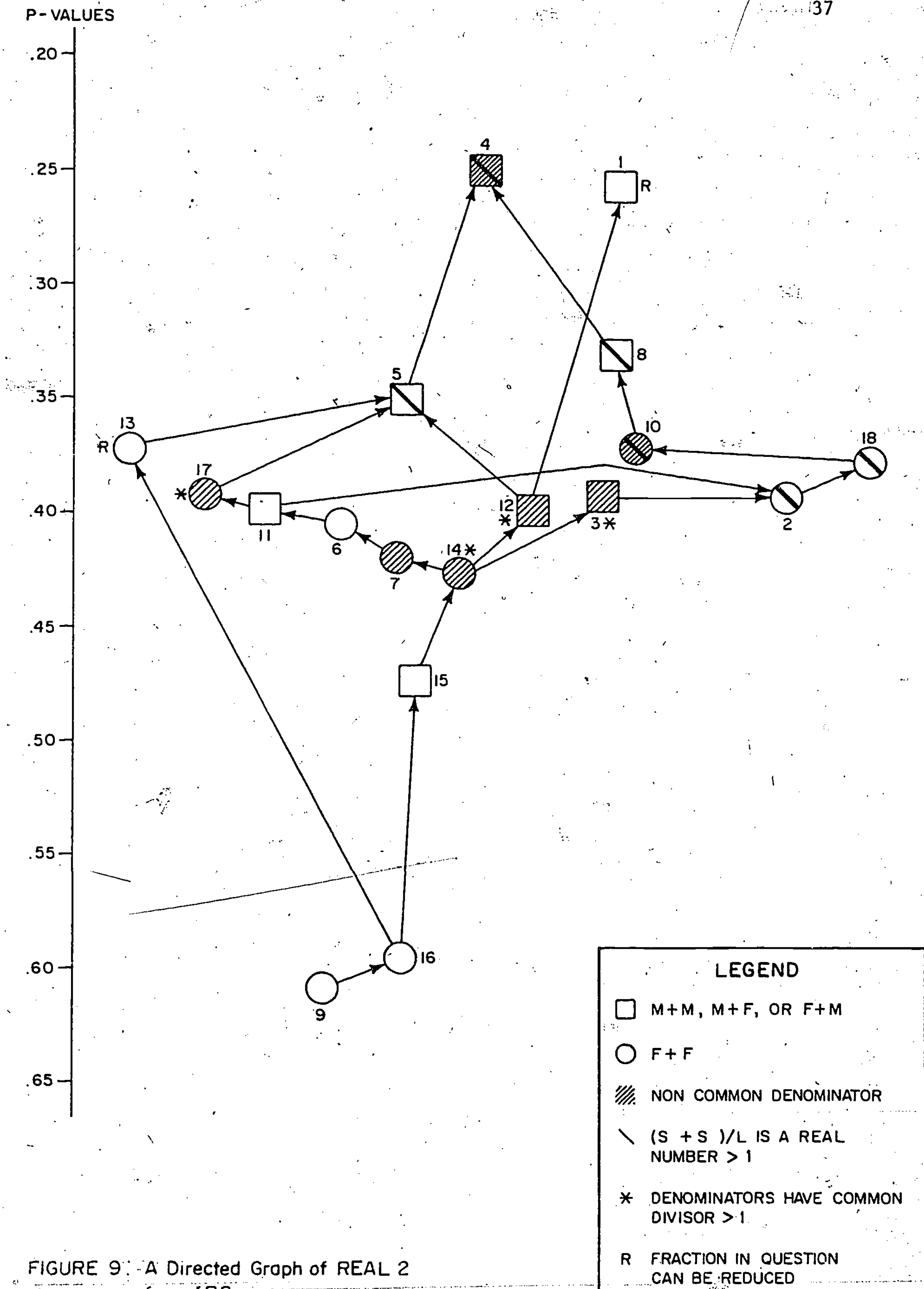


FIGURE 9: A Directed Graph of REAL 2 from IRS

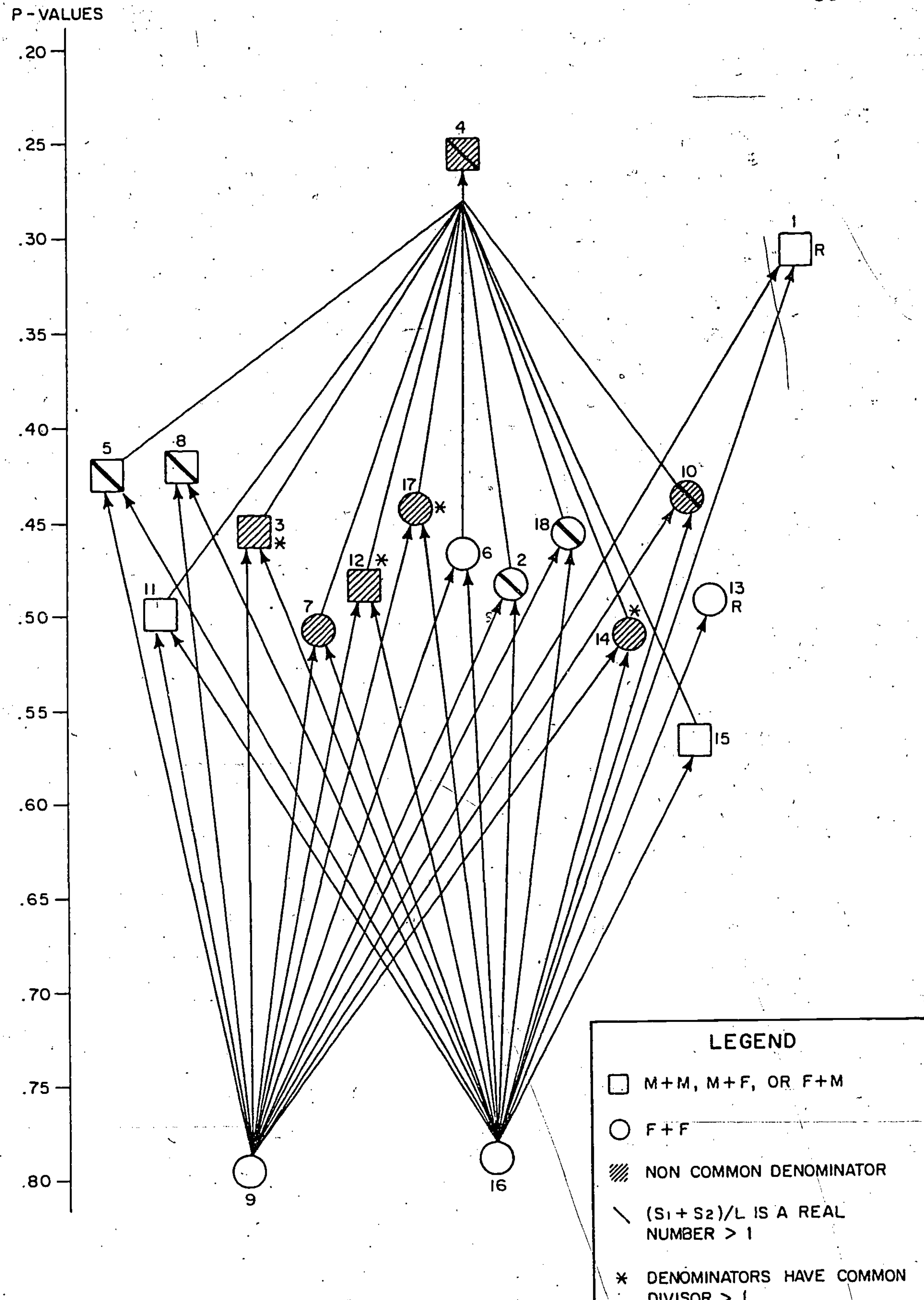


FIGURE 10. A Directed Graph of NSIM from ORDER 2

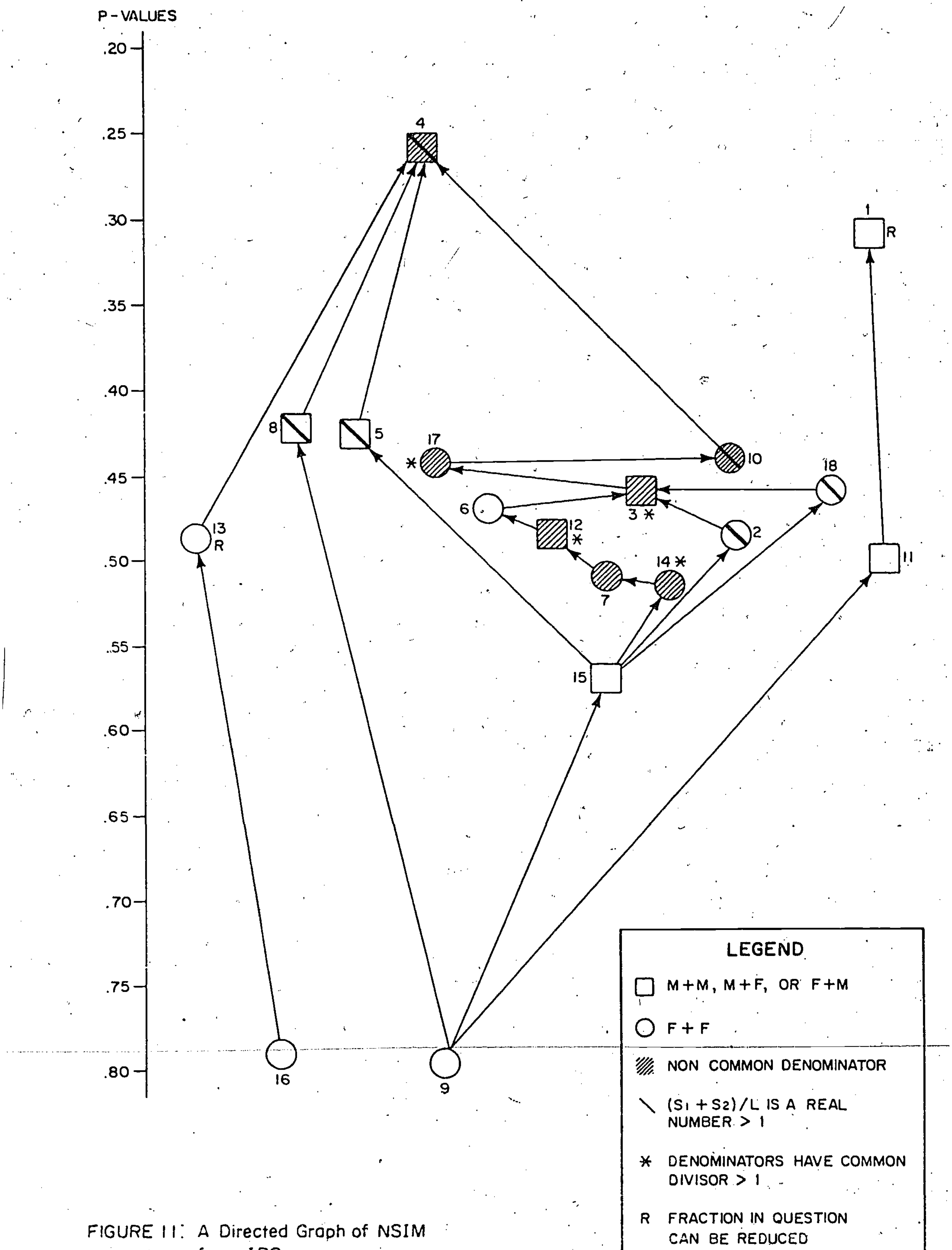


FIGURE 11: A Directed Graph of NSIM from IRS

item means may, however, be caused by the distributions of θ. Close analysis of the properties of any simulated dataset is required before it is employed in any study.

Summary and Discussion

Two order analytic approaches to the analysis of test structure have been presented and described. It was shown that for unidimensional data the Krus and Bart procedure more closely reconstructed the procedural network for fraction addition than the procedure proposed by Tatsuoka and Tatsuoka based on Takaya's IRS matrix. Thus, when trying to discover the relationships of procedural skills and to sequence instruction accordingly, this procedure supplies more information about the hierarchical structure of tasks. Use of IRS though, appears to be more appropriate when large amounts of error may be in the data. This is apparent from its ability to extract a structure from simulated data.

Clearly, caution is warranted in the use of simulated data in quantitative research of the type carried out in this study. It was shown that not only can the means, variances, a, and b values, of simulated datasets be greatly affected by the "random" nature of the simulation procedure and the original a and b values used as input but that the hierarchical structure of the data is also greatly altered. The currently used simulation technique is inadequate in reproducing the data when a set of a values which include exaggerated a's is used as the basis of the simulation. Furthermore, it was shown that this simulation technique can greatly alter the item difficulties. This may be due to the fact that the distribution of ability is not accounted for in the population.

Research in this area should include a large scale sampling distribution study to determine the distribution properties of simulated data. A more sophisticated method of generating binary responses which accounts for the destribtuion of θ needs to be developed. Also, a significance test and possibly a test of the differences between two item characteristic curves should be investigated.

References

Antonak, R., Bart, W., & Lele, K. ORDER2: A computer program to perform ordering-theoretic data analysis, 1979.

Airasian, P. W., & Bart, W. Ordering theory: A new and useful measurement model. Educational Technology, 1973, 13, 56-60.

Airasian, P. W., & Bart W. Validating a priori instructional hierarchies. Journal of Educational Measurement, 1975, 12(3), 163-173.

Airasian, P. W., Madaus, G., & Woods, E. Scaling attitude items: A comparason of scalogram analysis and ordering theory. Educational and Psychological Measurement, 1975, 35(4), 809-820.

Baillie, R., & Tatsuoka, K. K. IRS: A computer program for extracting item hierarchies based on a modified Takeya's IRS analysis on the PLATO[R] system, 1981.

Baillie, R. GETAB: A computer program for estimating item and person parameter of the one- and two-parameter logistic model on the PLATO[R] system, 1982.

Bart, W. An empirical inquiry into the relationship between test factor structure and test hierarchical structure. Applied Psychological Measurement, 1978, 2(3), 331-335.

Bart, W., & Krus. D. An ordering theoretic method to determine hierarchies among items. Educational and Psychological Measurement, 1973, 33, 291-300.

Guttman, L. L. Studies in social psychology in world war II. In S. A Stouffer (Ed.), Measurement and Prediction (Vol. 4). Princeton, N.J.: Princeton University Press, 1950.

Horst, P. Correcting the Kuder-Richardson reliability for dispersion of item difficulties. Psychological Bulletin, 1953, 50, 371-374.

Klein, M. F., Birenbaum, M., Standiford, S. N., & Tatsuoka, K. K. Logical error analysis and construction of tests to diagnose student "bugs" with addition and subtraction of fractions (Research Report 81-6-NIE). Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, November, 1981.

Krus, D. Logical basis of dimensionality. Applied Psychological Measurement, 1978, 2(3), 321-329.

Krus, D., & Bart, W. M. An ordering theoretic method of multidimensional scaling of items. Educational and Psychological Measurement, 34, 525-535.

Krus, D., & Krus, P. Dimensionality of hierarchical and proximal data structures. Applied Psychological Measurement, 1980, 4(3), 313-321.

Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. Psychological Bulletin, 1947, 45, 507-529.

Lord, F. M. Applications of item response theory to practical testing Problems. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

McNemar, Q. Notes on sampling error of the differences between correlated properties of percentages. Psychometrica, 1947, 12(2), 153-157.

Sato, T. Personal communication, July 4, 1981.

Shevell, S. K. A scalability coefficient for dominance and proximity data (Research Report). Ann Arbor, Mich.: University of Michigan, Department of Psychology, December, 1975.

Takeya, M. Relational structure analysis among test items on performance scores. Journal of Science Education in Japan, 1980a, 4(4), 183-192. (in Japanese)

Takeya, M. A method of structuring IRS graphs and its application. Japanese Journal of Educational Technology, 1980b, 1(5), 93-103. (in Japanese)

Takeya, M. A study on item relational structure analysis of criterion reference tests. Unpublished doctoral dissertation. Waseda University, Tokya, 1981. (in Japanese)

Tatsuoka, K. K., & Chevalaz, G. A map representation of misconceptions: an approach utilizing item response theory and classification functions (Research Report 82-5-ONR). Urbana, Ill.: University of Illinois, Computer based Education Research Laboratory, 1982.

Tatsuoka, K. K., & Tatsuoka, M. M. Item analysis of tests designed for diagnosin bugs: Item relational structure analysis method (Research Report 81-7-NIE). Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, 1981.

Tatsuoka, M. M. Personal communication, July 23, 1981.

Wise, S. L. A modified order analysis procedure for determining unidimensional item sets. Unpublished doctoral dissertation, University of Illinois, 1981.

Sato, T. Personal communication, July 4, 1981.

Shevell, S. K. A scalability coefficient for dominance and proximity data (Research Report). Ann Arbor, Mich.: University of Michigan, Department