

DOCUMENT RESUME

ED 395 312

CS 215 277

AUTHOR Burley, Hansel; And Others
 TITLE A CHAID Analysis of a Diagnostic Writing Sample as a Placement Tool for Freshman Composition.
 PUB DATE 27 Jan 96
 NOTE 34p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 25-27, 1996).
 PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
 EDRS PRICE MF01/PC02 Plus Postage.
 DESCRIPTORS Basic Writing; Community Colleges; *Freshman Composition; *Grading; Higher Education; Open Enrollment; Statistical Analysis; *Student Evaluation; *Student Placement; *Writing Evaluation; Writing Research; *Writing Skills; Writing Tests
 IDENTIFIERS Chi Square Automatic Interaction Detector; *Placement Tests; Texas; Writing Samples

ABSTRACT

Proper placement of students into either remedial writing or Composition I can be crucial to their success in higher education. Using a database of nearly 6,000 students who entered an open admissions community college in Texas, the researchers attempted to discover the best predictor of student success in Composition I. For students who took a locally scored entry/placement test, the best predictor of success in Composition I was the reading portion of the test, not the writing portion. For students taking a statewide test, the best predictor was the writing portion of that test. For students taking both the local and the statewide test, the best predictor of Composition I performance was passing or failing the writing portion of the statewide test. The researchers concluded that the major differences between the 2 groups had to be the grading practices of the locally administered test. The researchers recommend that the community college emulate the grading procedures used by the administrators of the statewide test and/or writing assessment theorists, like E. M. White, who maintain that holistic scoring, with all its notations, is the most successful method of scoring writing in quantity that is now available. He recommends the use of 6 procedures: (1) controlled essay reading; (2) a scoring criteria guide; (3) anchor papers; (4) checks of reading in progress; (5) multiple independent scoring; and (6) evaluation and record keeping. (Contains 3 tables of data and 3 figures.) (Author/TB)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED 395 312

A CHAID Analysis of a Diagnostic Writing Sample
As a Placement Tool for Freshman Composition
(Alternative title: The Importance of Proper Procedure in the Scoring of Diagnostic Essays)

by

Hansel Burley, David England, and Paul Beran

Paper Presented At the

Southwest Educational Research Association
January 27, 1996

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

H. Burley

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U. S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as
received from the person or organization
originating it.

Minor changes have been made to
improve reproduction quality.

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

BEST COPY AVAILABLE

2

CS 2/5 277

A CHAID Analysis of a Diagnostic Writing Sample
As a Placement Tool for Freshman Composition
(Alternative title: The Importance of Proper Procedure in the Scoring of Diagnostic Essays)
by

Hansel Burley, David England, and Paul Beran

Paper Presented At the

Southwest Educational Research Association
January 27, 1996

ABSTRACT

Proper placement of students into either remedial writing or Composition I can be crucial to their success in higher education. Using a database of nearly 6,000 students who entered an open admissions community college, the researchers attempted to discover the best predictor of student success in Composition I. For students who took a locally scored entry/placement test, the best predictor of success in Composition I was the reading portion of the this test, not the writing portion. For students taking a statewide test, the best predictor was the writing portion of that test. For students taking both the local test and the statewide test, the best predictor of Composition I performance was passing or failing the writing portion of the statewide test. The researchers concluded that the major differences between the two groups had to be the grading practices of the locally administrated test. The researchers recommended that the community college emulate the grading procedures used by the administrators of the statewide test and/or writing assessment theorists like E. M. White.

A CHAID Analysis of a Diagnostic Writing Sample
As a Placement Tool for Freshman Composition
(Alternative title: The Importance of Proper Procedure in the Scoring of Diagnostic Essays)

Hansel Burley, David England, and Paul Beran

It goes without saying that the correct placement of entering freshmen into either Composition I or developmental writing is crucial to students' academic success. Perhaps, however, it has not been said enough. When one adds emotional and financial variables of being placed in a noncredit course, especially for minority students, first time in college students, and students from low socioeconomic backgrounds, correct placement is a moral act, a concrete illustration of a college's mission in action. Haphazard placement practices actually backfire; nothing could turn a student away from a school faster than unreliable advisement and placement. To make matters worse, this process can be distorted even more by the use of the holistic scoring of essays, administered without scoring sessions grounded in tried and tested procedures. Some misplaced students may turn away from the institution; those improperly placed in remediation may accept the institution's label and behave like remedial writers; those improperly placed in Composition I may be emotionally crushed by the succession of seemingly inexplicable F's they receive.

Therefore, as English teachers, writing assessment administrators, and educational researchers, we must strive to ensure a test's construct validity and its consequential validity. By construct validity, we mean that the test measures what it purports to measure (Borg & Gall, 1989). Writing tests ought to be constructed, delivered, and scored so that they accurately

measure writing ability. Consequential validity involves a concern for the intended and unintended consequences of an assessment (Cronbach, 1988). Rightly so, composition instructors demand that actual writing be a part of any entry assessment; when scoring these papers, then, it is essential that we get it right.

Clearly, as E. M. White says, "holistic scoring, with all its notations, [is] . . . the most successful method of scoring writing in quantity that is now available" (1988 p. 30). He continues, "This method of scoring has made the direct testing of writing practical and reliable: it indirectly and effectively brings together English teachers to consider and discuss the goals of writing instruction and it embodies a concept of writing that is responsible in the widest sense" (p. 30). He states the following:

Holistic scoring is able to achieve acceptably high reliability by adding a series of constraints to the economically efficient practice of general impression scoring. Basic to all these constraints is a carefully developed and precise writing assignment (sometimes called a "prompt"), followed by an attempt to reduce unnecessary variability in the scoring process. Six procedures and practices have been developed for scoring, and where all six are observed with sensitivity and care, high reliability of scoring has been achieved with no appreciable sacrifice of economy. (p. 23-24)

Those six procedures are (1) controlled essay reading, (2) scoring criteria guide, (3) anchor papers, (4) checks of the reading in progress, (5) multiple independent scoring, and (6) evaluation and record keeping. The limitations of holistic scoring include (a) little meaningful diagnostic information beyond rank ordering of papers, and (b) reliabilities can be overestimated (White

1988). English departments, well intentioned though they may be, that do not understand how easy it is to undermine correct holistic scoring of diagnostic essays and hence the correct placement of students, can be the most recalcitrant of offenders. However, when one weighs the needs of the learners, especially the needs of vast numbers of students who potentially may be misplaced, professionalism and goodwill demand that scorers and leaders of readings at least change their behavior, if not their attitudes about the placement of students into freshman composition classes.

With the above in mind, the purpose of this study was to check the effectiveness of a community college's placement practices for freshman composition and developmental writing courses. Specifically the research questions is

Will the college's locally administered composition placement test adequately predict student scores in Freshman Composition I as well as the state administered tests?

Ideally, what one would want from a placement test is strong predictive validity--that is a high positive correlation between scores on the placement test and performance in the freshman composition course. According to Borg and Gall (1989) "predictive validity is the degree to which the predictions made by a test are confirmed by later behavior of the subjects" (p.252). They continue, noting the importance of carefully choosing criterion measures (placement test scores and course grades in this case) and "*measurement procedures used to obtain*" [italics added] those scores (p.253).

Finally they stress that "It is important to assess the predictive validity of a standardized test before deciding whether to use it in making practical decisions requiring forecasts, such as selecting students for colleges" (p. 253). Strong predictive validity of a diagnostic writing sample

is important for accurate placement decisions.

Typically, the first year college writing program has entry at several levels, possibly ranging from a grammar and paragraph writing course, to an accelerated honors writing course. These tests, then, help institutions to arrange students in efficient, homogenous teaching groups. This activity is not necessarily associated with the curriculum, so the use of placement tests is controversial (White, 1994). Oftentimes these assessments are quite unlike assesment in the writing classroom, where a student's behavior may be characterized by fits and starts sprinkled with refreshing insights. If writing as a recursive process is the norm, how can a sample of student work, written in 90 minutes accurately represent a student's ability? On the other hand, these placement tests provide needed information about a student--fast. This information can be used to establish accountability systems that satisfy researchers, administrators, and legislators; and in this time of shrinking budgets, all involved want to know what they are getting for their money. Are the students really learning anything in that composition course? Will performance on the test predict future behavior?

Method

We paired freshman composition grades with placement test data. Since this Texas community college uses basically two entrance tests (a local assessment and a statewide test), the data were divided into two groups, students having only the statewide test recorded on their records, and those having the local placement test on their records. Some students may have taken both exams. A special analysis was done of these students.

The Texas Academic Skills Program (TASP) provides information about the reading, mathematics, and writing skills of students entering Texas public universities and colleges.

Based on student performance on this test, universities and colleges are required to provide support services and remedial courses and activities for students who failed to pass one or more sections (reading, mathematics, and writing) of the TASP test. These students must remain in continuous remediation until they pass all sections of the test. Both the Pre-TASP and the TASP draw questions from the same question bank, but the Pre-TASP is half as long. Whereas Pre-TASP sessions last from two to three hours (varies from institution to institution), TASP sessions last for five hours. Both the Pre-TASP and the TASP are developed by National Evaluation Systems. These are the tests used in this study.

The statistical procedure used was CHAID (Chi-square Automatic Interaction Detector). CHAID divides a population into distinct groups based on categories of the best predictor of the dependent variable. In this study, the dependent variable was Composition I failure or success. The predictor variables were students' failure or passing of the reading, composition, and mathematics portions of the Pre-TASP or TASP tests. The CHAID analysis then further re-analyzes these subgroups of the best predictor based on other predictor variables. Researchers usually present CHAID analyses as a tree, with each branch indicating a statistically significant interaction and each level indicating newly formed subgroups. Principally, this study is interested only in the best predictor for the dependent variable, freshman composition final grade. Failure was any grade other than A, B or C. Therefore, D's, F's, W's and I's were categorized as failure in the course. W's denote withdrawals, and I's represent incompletes.

Limitations

These are naturally occurring groups, so there is no random assignment to treatment (Pre-TASP group and TASP group) in this study. Since the Pre-TASP is a free test and since TASP

costs twenty-six dollars, poorer students may have self-selected themselves into the Pre-TASP group. The research hope, however, is that the size of the database used in this study may ameliorate some of the effects of this possible self-selection.

Analysis

Descriptives

The entire study had a total of 6,338 students. The majority were white students of European descent, nearly 75% of the population. African-Americans made up almost 14% of the students taking either the Pre-TASP or the TASP while Hispanic students were nearly 10% of the population (See Table 1). The study drew upon data from nine semesters (See Table 2). When this population was split into those placed into freshman composition because of Pre-TASP scores or TASP scores, these demographics remained fairly stable for the 5,149 remaining students (See Table 3). Fifty-eight percent of these students in the study were women and almost 42% were men.

Table 1
Gender and Ethnic Data for All Students in Stud

	European	Afri-Am.	Hispanic	Asian	Nat. Am.	Non Res.	Row Totals
Women	2463	601	336	29	1.4	1.4	3637
Row %	72.7	16.5	9.2	.8	.4	.4	57.4
Column %	55.7	68.5	57.0	49.2	70.0	31.8	
Men	2105	277	253	30	6	30	2701
Row %	77.9	10.3	9.4	1.1	.2	1.1	42.6
Column %	44.3	31.5	43.0	50.8	30.0	68.2	
Column	4748	878	589	59	20	44	6338
Total%	74.9	13.9	9.3	.9	.3	.7	100

Table 2
Number of Students in Study By Semester

Semester	Number	Percentage
1990--Fall	818	12.9
1990--Spring	501	7.9
1991 --Fall	872	13.8
1991 --Spring	541	8.5
1992--Fall	946	14.9
1992--Spring	538	8.5
1993--Fall	956	15.1
1993--Spring	622	9.8
1994--Spring	544	8.6
Total	6338	

Table 3
Success of Study Population in Freshman Composition I

	European-Am.		Afri-Am.		Hispanic		Totals+
	Passed	Failed	Passed	Failed	Passed	Failed	
Women	1504	708	227	207	148	139	2978
Men	958	759	62	134	79	125	2171
Total Pass	2462	1467	269	361	227	264	

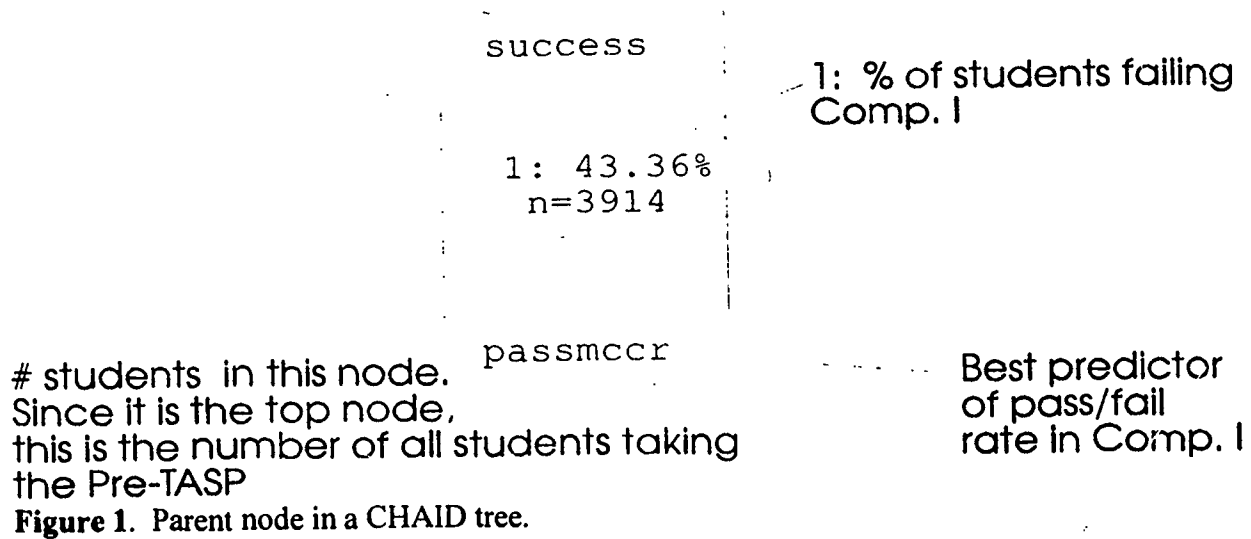
In general, European-Americans out perform African-Americans and Hispanics in freshmen composition, and women outperform men. Asians, Native Americans, and

Nonresident aliens are only 2% of those students who had either TASP or Pre-TASP scores, so they are not presented in Table 3. However, these students are included in any analysis.

CHAID Analysis

As mentioned above, the data were divided into two groups--those students placed by the Pre-TASP and those placed by the TASP. For those students who took the Pre-TASP, surprisingly the best predictor for success in freshman composition was student performance on the reading section of the test, not the writing section (See Figure 1).

Outcome variable: Pass/fail Comp. I



Since CHAID is a new statistical procedure, an explanation of how to read a CHAID tree will accompany the discussion. Significant interactions in a CHAID analysis are placed in rectangles called nodes. Figure 1 is an example of only one part of a CHAID tree, the parent node; see the back of this paper for the full tree. At the top of the node in Figure 1 is the word "success." This is the name of the outcome variable of interest, success or failure in

Composition I. The "1" in this node refers to the levels of the success variable; in this case there are only two, and the "1" represents the percentage of students failing freshman composition. The "n=3914" refers to all students in the analysis for this node. Since this is the top or parent node, this is the number of all students taking the Pre-TASP. We entered three predictor variables in the model, pass/fail Pre-TASP writing, reading, and math. CHAID places the best predictor variable, the one with the closest association to passing Composition I underneath this parent node. "Passmccr" is the name we gave to the reading test. Therefore, performance on the reading test is the best predictor of the three entered in the model. In other words, in comparison with the writing and math batteries, failure on the reading test is the best predictor of failure in Composition I, and passing the reading test is the best predictor of passing Composition I. Finally, CHAID then reanalyzes the reading score (the best predictor) for a significant interaction and finds it (See Figure 2 at the end of this paper). The reading variable has two levels, pass and fail, and these two groups are significantly different from each other. Reading from left to right, of those who failed the reading exam, 1,323 students, 50% failed Composition I. One should compare this to node to the one on the right which represents students who passed the reading exam. In this case, of the 2,591 students who passed the reading test, only 40% failed Composition I. On the last level of the CHAID tree, is node -3-; here are the students who passed the reading test, but failed the writing test. Forty-four percent of these students failed Composition I. Compare this number to the students that both passed the reading and writing sections of the Pre-TASP. Thirty-eight percent of these students failed Composition I.

This tree directly informed our recommendations to the English department at this Texas community college. Because the most accurate predictions of student success in Composition I

include passing the reading entry test, reading scores should be used to help place students, especially those students making borderline pass/fail writing placement test scores. More importantly, this tree dramatized the need to revisit and reorder diagnostic scoring procedures, with the aim of making the writing test a more reliable predictor of student behavior.

Figure 3 is a CHAID analysis of those students who took the TASP. This is a statewide test that has a highly developed writing test scoring procedure that is heavily dependent on establishing high inter-rater reliabilities between scorers. For the students taking this test, the best predictor was the writing test--as it should be. As a secondary analysis, we studied those students who had taken both Pre-TASP and the TASP tests (See Figure 4). Six variables, reading, writing, and math performance on both tests were entered into the model. Here again, the best predictor of success is performance on the TASP writing test (paswrit1). Finally, several subgroup analyses, not featured in this report, confirmed the above analysis in all but one case. In these analyses, the students were divided by the semester they enrolled into Composition I and when they took the PreTASP and TASP. We selected for the smallest time differences between these events. In the one case not confirmatory of the above analysis, the best predictor was the Pre-TASP reading (Fall 1993 students). We are certain that this is influenced by a higher "cut-off" score on the Pre-TASP reading test than on the TASP reading test.

Discussion/Implications for Practice

Will the college's locally administered composition placement test adequately predict student scores in Freshman Composition I as well as the state administered tests? It is surprising that the best predictor for success in Freshman Composition I is the reading section of the Pre-

TASP and not the writing section. Confounding matters more is the fact that the best predictor for success on the TASP test was the writing section of the test. These tests should be highly correlated, but at least the writing sections do not seem to be.

There seem to be three contributory causes to this phenomenon: 1) Two different types of students take the two tests, 2) the grading and administration of the tests are different, 3) some combination of the two. Causal agent two is probably the most important, and fortunately, it is one that the college can do something about. Students who take the TASP have to pay a fee to take it, whereas the Pre-TASP is free; therefore, students with a higher socioeconomic status, higher self-efficacy and motivation, and/or higher aptitudes or previous achievement may tend to take the TASP. So, the TASP takers may be better writers at the start, and the correlation between their writing test scores and Composition I grades ought to be high. Just the opposite could be true for the students who take the Pre-TASP. Still, the Pre-TASP writing test ought to predict the behavior of students who perform less well on the test and are less prepared just as accurately as the TASP.

Hence, the grading and administration of the test may be the problem. For the grading of the statewide TASP, inter-rater correlations were run to ensure that graders are grading consistently, and the TASP administrators adhere to the principles of holistic grading religiously.

At that Texas community college, the grading of the Pre-TASP fell short of the TASP administration standards. In one case, one grader gave 3's (on a scale of 1 to 4) to every single paper in a group of 60 papers. This person helped grade hundreds of papers. Standard operating procedure for the scoring of the Pre-TASP papers was for the testing administrator to leave the papers in an instructor's campus mailbox with note saying that the papers would be needed back

in two hours. Sometimes graders were literally "captured" as they walked down the hall and pressed into service by the writing test administrator. Calibrations were rare, and inter-rater reliabilities were nonexistent. Certainly, one thing remains clear, the writing section of the Pre-TASP, the locally scored assessment, ought to be the best predictor of success in Freshman Composition 1; and for too many students, it was not.

Suggestions

- The English department at this college should re-evaluate current practices, taking pains emulate the grading procedures of the TASP and those tried and tested procedures promulgated by E. M. White.
- Further study is needed of the characteristics of the students who take the TASP and those who take the locally graded assessment. For example, does gender or ethnicity play a role?
- A certain cut score on the reading test may need to be added as part of the criteria for entering a particular English composition course. For certain students, reading may need to be a prerequisite for entering freshman composition.

Final Note

Based upon the results of this study, the English department agreed to add performance on the reading test as a tertiary criterion for those troublesome borderline- pass-or- fail writing samples (primary criterion was the writing sample; the secondary criterion was an objective multiple choice grammar and usage test). This department is even considering making reading a prerequisite to Composition I for those borderline students, especially since the reading

objectives on the test are similar to the reading objectives in Composition I. More importantly, the English department adopted more rigorous calibration and scoring procedures, wisely using E. M. White's (1988) six procedures.

References

Borg, W & Gall, M. (1989). *Educational research. An Introduction* (5th ed.). New York:

Longman

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer (Ed.). *Test validity*

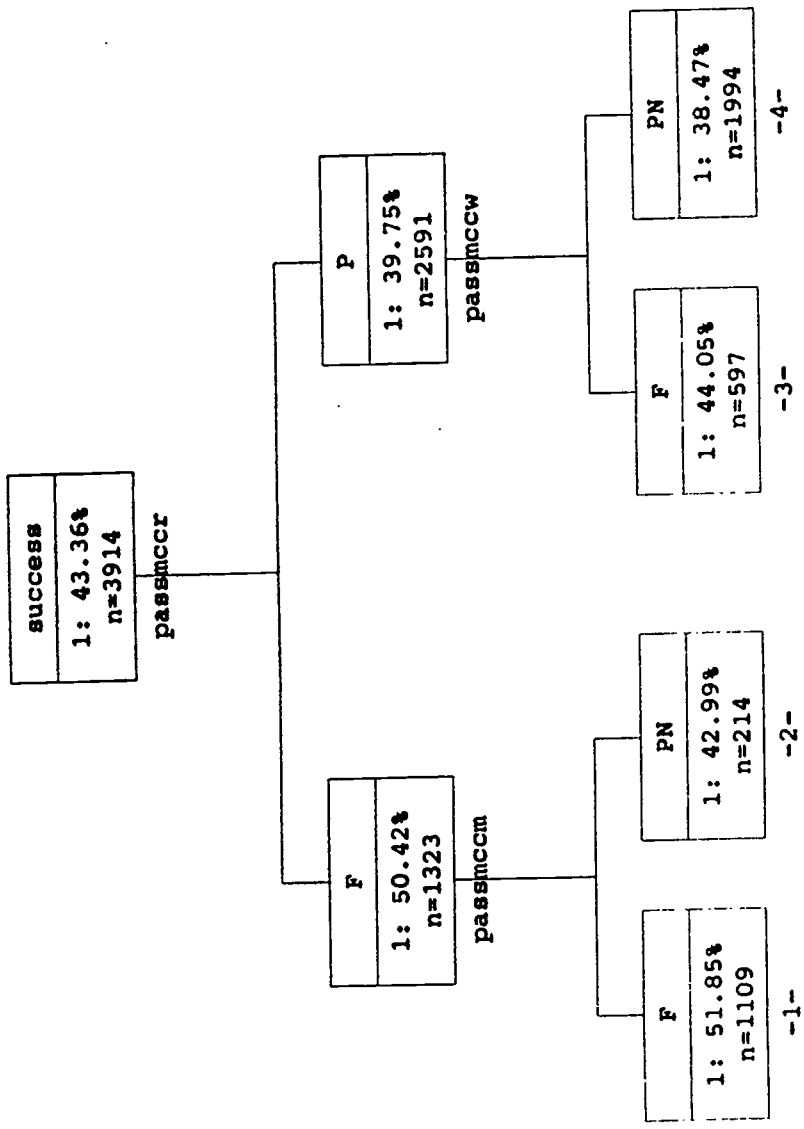
(pp.3-17). Hills dale, NJ: Erlbaum.

White, E. M. (1988). *Teaching and assessing writing*. San Francisco: Longman.

White, E. M. (1994). Issues and problems in writing assessment. *Assessing Writing* , 1 (1), 11-

27.

Figure 2. CHAID tree for students taking the Pre-TASP. Passmccr or the reading section of the test is the best predictor for success or failure in Composition I.



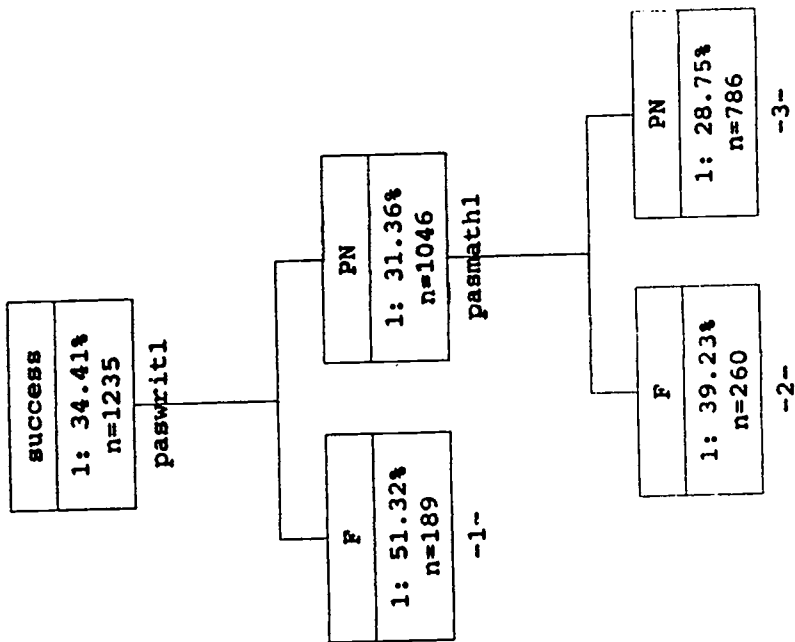
HAIID Tree for Students Taking the Pre-TASP

Table of passmccr by success

passmccr (before)	not success		row % Total
Failed	50.42	49.58	1323
Passed	39.75	60.25	2591
Total	43.36	56.64	3914

LR chi-square=40.40 df=1 prob=2.1e-10

Figure 3. CHAID tree for students taking the TASP. Paswrit1 or the writing section of the TASP is the best predictor of student success or failure in Composition I.



HAIJ Tree for Students Taking the TASP

Table of paswrit1 by success

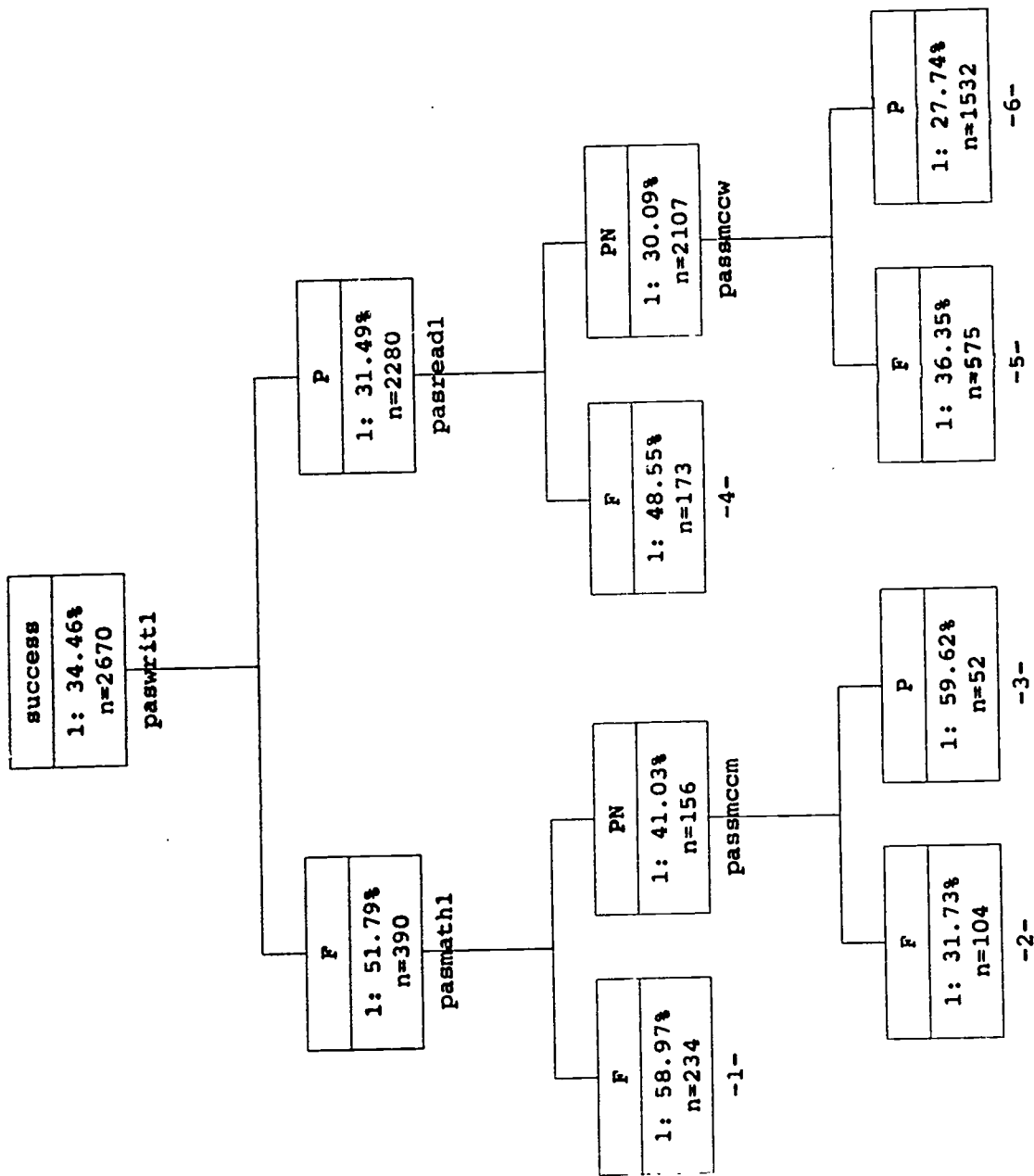
paswrit1 (before)	not success	row % Total
Failed	51.32 48.68	189
Passed	31.16 68.84	1027
Not Taken	42.11 57.89	19
Total	34.41 65.59	1235

LR chi-square=28.06 df=2 prob=8.1e-7

20

20

Figure 4. CHAID tree for students taking both the Pre-TASP and the TASP. Paswrit1 or the writing section of the TASP is the best predictor of student performance in Composition I. Six variables, reading, writing, and math from both test, were used for this analysis.



20

30

AID Tree for students Taking Both Pre-TASP and TASP
Table of paswrit1 by success

paswrit1 (before)	not success	row % Total
Failed	51.79 48.21	390
Passed	31.49 68.51	2280
Total	34.46 65.54	2670

LR chi-square=58.15 df=1 prob=2.4e-14

HAIID Tree for students Taking Both Pre-TASP and TASP

Table of paswrit1 by success

paswrit1 (before)	not success	row % Total
Failed	51.79 48.21	390
Passed	31.49 68.51	2280
Total	34.46 65.54	2670

LR chi-square=58.15 df=1 prob=2.4e-14