DOCUMENT RESUME

ED 317 600 TM 014 695

AUTHOR Micceri, Theodore; And Others
TITLE Consistent Patterns in Observed Teacher Performance: Results from a Large-Sample Multi-Year Study. Draft.
PUB DATE 11 Mar 90
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; Educational Improvement; *Elementary School Teachers; Elementary Secondary Education; Longitudinal Studies; Problem Solving; Questioning Techniques; *Secondary School Teachers; Summative Evaluation; Teacher Behavior; *Teacher Evaluation; Teaching Methods
IDENTIFIERS Florida Performance Measurement System; *Performance Indicators

ABSTRACT

Observation of teacher classroom performance can provide insights into the steps necessary to improve pedagogical practice. Data from over 13,000 field observations of Florida teachers over a 2-year period revealed parallel performance patterns across grade levels and years among 40 behavioral indicators. The study used the Florida Performance Measurement System summative observation instrument--a measure of teacher classroom behavior that is comprised of 20 behavioral indicators shown to be effective and 20 indicators shown to be ineffective by the educational process-product literature. Data from 7,926 observations (4,447 at the elementary school level and 3,479 at the secondary school level) conducted during the 1987-88 school year and from 4,575 observations (2,735 at the elementary school level and 1,840 at the secondary school level) conducted during the 1988-89 school year were submitted to analysis. Questioning and interactive indicators, both effective and ineffective, dominated the average lesson (60%), while indicators dealing with higher order thinking and problem solving comprised only 4% of all behaviors. The data imply that awareness of ineffective questioning behaviors may need emphasis in teacher training programs and that the current emphasis on problem solving is appropriate. Twelve graphs and one table are provided. (Author/TJH)

# Consistent Patterns in Observed Teacher Performance: Results from a Large-Sample Multi-Year Study



by
Micceri, T.
Peterson, D.
Borg, J. M.



Paper presented at AERA annual conference, Boston, MA, April 16-20, 1990

## Abstract

Observation of teacher classroom performance can provide insights into the steps necessary to improve pedagogical practice. Data from over 13,000 field observations of Florida teachers over a two year period revealed parallel performance patterns across grade levels and years among 40 behavioral indicators. Questioning and interactive indicators, both effective and ineffective, dominated the average lesson ( 60%) while indicators dealing with higher order thinking and problem solving comprised only four percent of all behaviors. These data imply that awareness of ineffective questioning behaviors may need emphasis in teacher training programs and that the current emphasis on problem solving is appropriate.

## Background and Development

Observation of teacher classroom performance can provide insights into the steps necessary to improve pedagogical practice. The Florida Performance Measurement System (FPMS) summative observation instrument is a measure of teacher classroom behavior comprised of 20 behavioral indicators shown to be effective and 20 indicators shown to be ineffective by the educational process-product literature (A copy of the summative instrument is contained in Smith, Peterson & Micceri, 1987). Development of the Florida Performance Measurement System (FPMS) began in 1979. The purposes of this research was to organize the substantial number of teacher effectiveness studies conducted over the past thirty years into a useable form for teacher training, and to create a system of evaluation based upon these researches that could be used in the assessment of teacher performance for both formative and summative purposes. This was timely in view of Florida legislation establishing a Beginning Teacher Program (BTP) requiring school districts to assist new teachers in their first year of instruction and to verify

the competence of newly graduated teachers prior to their receiving a state teaching certificate.

The first phase of development involved assembling the research on generic teaching modes produced by process-product and experimental studies into an organized knowledge base. Six domains of teacher performance indicators resulted: planning, management of student conduct, organization and development of instruction, presentation of subject matter, verbal and nonverbal communication, and student testing. The domains included 34 major instructional concepts and 124 indicators of effective or ineffective teaching behaviors. The basic content validation of these concepts and indicators derives from historic research showing each to be either effective or ineffective in classroom situations. Formal content validation utilizing widely recognized experts in the field was conducted in 1983.[1]

Once the research was organized, it proved feasible to develop formative observation instruments for the four observable domains (other than planning and testing) and a cumulative summative instrument incorporating synthesized indicators from the formative instruments. The summative instrument was designed to serve two functions: 1) as a screening measure, it enables a trained observer to identify a specific teacher's problem areas (e.g. management of student conduct or organization and conduct of instruction) and then to utilize more specific formative instruments to pinpoint specific teaching behaviors that require remediation, and 2) the instrument can also be used for final evaluations.

An instrument designed to measure and evaluate the effectiveness of teaching performance must be reliable, valid and normed. Several studies have been conducted on the FPMS summative instrument to evaluate these characteristics. Resources were concentrated on the summative instrument due to professional and legal considerations resulting from its use for decision-making purposes. An extensive G-study was conducted to determine the reliability of this instrument. The coefficients obtained from this study for teams of two observers were: intercoder agreement r=.85, stability, r=.86 and reliability, r=.79.[2] Since this study, several field studies have produced similar results, and one major study on stability produced a

---

[1] *Teacher Evaluation Project: Final Report for 1982-1983* (Document No. SP 028190), Tampa: University of South Florida (ERIC Document Reproduction Service No. ED 266 121).

[2] *ibid.*

test-retest reliability coefficient of r=.70 (n=33, $\rho$<.001) with a six-month time lapse.[3] These results suggest that the behaviors being examined are extremely stable over time for experienced teachers.

A norming study was conducted on the FPMS summative instrument in 1983-1984.[4] Results of this study, based on field observations by certified observers of 1,223 teachers in grades K-12, found the instrument to in fact be generic. No differences in teacher scores could be attributed to sex, race, subject area, experience, class size, SES and other factors. There was a significant difference in scores between elementary and post-elementary grades. This difference was controlled by the creation of two norm groups (K-6 and 7-12). A difference based on instruction format (whether interactive or non-interactive) was controlled through the scoring procedure.

Five studies have been completed to estimate the predictive validity of the FPMS summative instrument.[5] Using the class as the unit of analysis, all five studies produced positive relationships between FPMS scores and student achievement measures. The conditional probability of this occurring randomly is ($\frac{1}{32}$ = .031). Meta-analytic tests of significance were significant at $\alpha$ < .01 for all five studies (132 classes, mean **r**=.29) and for three elementary studies ( 74 classes, mean **r** = .31); and were significant at the $\alpha$ < .05 level for the two mathematics studies (45 classes, mean **r** = .33), the biology study (20 classes, **r**=.43) and the two secondary studies (58 classes, mean **r**=.27). The only meta-analytic result not achieving significance was for the two social science classes, with **z** = 1.34 (67 classes, mean **r**=.18).

Florida provides an extensive training program for observers. A three-day period of observation training is followed by a criterion test of observation competence, an examination on teacher effectiveness research, a coding quiz and periodic update sessions. Over 9,200 trained and certified observers are currently active throughout Florida.

---

[3] *Assessing the Stability of the Florida Performance measurement System Summative Observation Instrument: A Field Study.* Miccori, T., Unpublished Technical Report, 1986.

[4] *Teacher Evaluation Study: Report for 1983-1984* (Document No. SP 027191), Tampa,: University of South Florida (ERIC Document Reproduction Service No. ED 266 122).

[5] *Florida Performance Measurement System Predictive Validity Report: A Meta-analysis of Five Predictive Validity Studies.* Unpublished technical report, Tampa: University of South Florida, 1987

A primary difference between the FPMS instruments and most teacher rating scales is its separation of observation and evaluation. Unlike rating scales that require the observer to code and evaluate simultaneously, the FPMS requires the observer only to code teacher behavior (independently of his or her opinion of its value). Evaluation is conducted at a later date through computer-based scoring of the observation instrument.

Adopted by 66 of 67 Florida school districts as the instrument of record for the Florida Master Teacher Program in 1984/1985, the instrument has been used in teacher performance evaluation by approximately 50 public and 20 private school districts in the state of Florida since the 1983/84 school year. In this way, the FPMS provides a means whereby the state Department of Education (DOE) as well as districts, colleges and individual schools and teachers can receive information on their relative performance. FPMS observations of a large sample of Florida beginning teachers have been recorded and stored in computer data bases at the Teacher Evaluation and Assessment Center (TEAC) in the University of South Florida's (USF) College of Education since 1984. The current study identifies behavioral patterns that occur among 38 of these 40 indicators of teacher performance contained in the FPMS summative instrument. One effective indicator Begins Promptly, and its ineffective counterpart, Delays, were not evaluated because each occurs or fails to occur only once in a lesson.

## Methods

Data from 7926 observations conducted during the 1987/88 school year (4447 elementary, 3479 secondary) and from 4575 observations conducted during the 1988/89 school year (2735 elementary, 1840 secondary) were submitted to analysis. All observations were conducted by certified observers. To be certified, observers must pass two tests of observational accuracy and an information test on the research literature underlying the instrument's items. Over 90 percent of the observations included in this study were conducted by either principals or assistant principals, with the remaining 10 percent conducted by peer-teachers, school district or university personnel. Although taking part in the BTP, many of the teachers included here (circa 45 percent) have prior experience, usually in another state. Over 50 percent of the teachers observed in the BTP recently moved to Florida from another state. This suggests that the findings of this study may be taken as representative of teachers in general, rather than only Florida beginning teachers.

Means of both scaled and raw item scores were computed separately for elementary (K-6) and secondary (7-12) teachers for the school years 1987/88 and 1988/89. Using the reported length of lessons, raw scores were standardized to a 30 minute lesson allowing an absolute comparison of performance frequency. Scaled scores allow for comparison against a normative group. The values use for scaling were derived from the FPMS Norming Study (Teacher Evaluation Study: Report for 1983-1994).[6] Based upon 1223 teacher observations, scaled values are assigned to each indicator independently by indicator coding frequency, Those scoring in the lower 25% (frequency) of norming study observations were assigned a 1, those between 26% and 75% a 2, and those greater than 75% a 3. For ineffective items, those in the upper quartile were assigned a 1 and those in the lower 75% a 2. Thus, for both effective and ineffective item scales, higher values represent respectively more effective behaviors or fewer ineffective behaviors. Therefore, each indicator is scaled positively, where a higher score is "better". The scaled effective and ineffective items were compared against each other to identify possible low scoring performance indicators. The sample $\bar{x}$ from the standardizing group is very close to 1.9 for each effective indicator. Since $s_x$ for scaled effective items consistently approximates .60, group mean scores falling between the value of 1.65 and 2.25 for effective items may be considered average relative to the performance of the large sample ($n > 1000$) of teachers from which these scalings were derived. Data were included from two different years to determine whether patterns of performance differ over time. Separate analysis of elementary and secondary observations also allows for the identification of specific similarities or differences across those teaching contexts.

This research sought to identify consistent patterns of behavior rather than test the significance of differences between individual items or groups. With samples of size 1000 or greater, $\bar{x}$ may be considered a good approximation of m, therefore means may be compared against standards such as the 1.65 to 2.25 range mentioned above. Due to the sheer volume of data involved, for reasons of space, analysis here is limited purely to comparative description.

---

[6] *op. cit.*

The FPMS behavioral indicators involved are listed in Figure 1. They fall into four domains of observable behavior one of which is broken into two subsets for better interpretation. Specifically, the domains are: Domain 3, Instructional Organization and Development; Domain 4, Presentation of Subject Matter; Domain 5, Communication: Verbal and Nonverbal, and Domain 2, Management of Student Conduct.

---

Insert Figure 1 About Here

---

## Results and Discussion

Raw Effective Items: Figure 2 shows mean raw effective item scores for a standardized 30 minute lesson. It is clear that mean scores for specific individual items were parallel for both years. Slight differences occur between elementary and secondary classes, with elementary teachers using more praise and practice as well as more behaviors in the domains of conduct and non-verbal communication than secondary classes. This was expected based upon all historic FPMS data, which clearly support the existence of two norm groups as reported above. By far the most frequently occurring items in both elementary and secondary classrooms for both years were Low Order Questions and their natural associate, Response/Feedback. Other frequently occurring items include High-Order Questions, Orienting/Focusing and, in the elementary classes, Stops Misconduct. In secondary classes, Emphasizing Important Points occurs more frequently than most items. The Items occurring an average of once or less during a 30 minute lesson include Beginning/Ending Reviews, Cause and Effects, Academic Rules and Criteria for Value Judgement.

Scaled Effective Items: Figure 3 shows the mean scaled item level performance scores derived from these data. The patterns were again the same from year to year, but slightly different between elementary and secondary classes. In general the most frequently occurring raw items represent higher than average (1.65-2.25) scaled scores. The highest scoring items in both elementary and secondary classes were. Orients/Focuses, the two Questioning items and Response/Feedback. In elementary classes,

Provides for Practice, Circulates/Assists, Positive Body Behavior and Stops Misconduct also showed comparatively high scaled scores. In secondary classes, Emphasizes Important Points and Treats Concepts exhibited fairly high mean scaled scores. The lowest scoring items for both elementary and secondary teachers occur in Domain 4, Presentation of Subject Matter (Cause and Effects, Academic Rules and Criteria for Value Judgements). Another item exhibiting low scaled scores was "Seatwork/homework". The only other item exhibiting a mean scaled score much below 2.0 is "Beginning/Ending Review".

---

Insert Figures 2 & 3 About Here

---

Raw Ineffective Items: Figure 4 shows the mean raw ineffective item scores for a standardized 30 minute lesson. Note here that the patterns are inverted from those of the scaled ineffective item scores in Figure 5. The scaling process for ineffective items is reversed from that used for effective items. Thus, more ineffective behaviors indicate "worse" performance. Scoring the two types of items (effective and ineffective) in this way allows for the computation of a cumulative total scaled score.

In elementary classes, General Praise occurs far more frequently than any other ineffective behavior[7], Allows Unison Response, and Multiple Questions Asked as One both average about one occurrence per 30 minute lesson.[8] In secondary classes, Allows Unison Response, Multiple Questions Asked as One and Non-Academic Questions occur with some frequency, as does, the dominant ineffective behavior, General Praise.

Scaled Ineffective Items: Figure 5 shows mean scaled ineffective items. For these scaled scores, as with the effective items, higher scores indicate "better" performance. For these scores, each hundredth point represents a percentage point. For example, a mean score of 1.90 means that 10 percent of all teachers exhibited that ineffective behavior. It is clear from this figure, that the same ineffective items occur most frequently both over years and across situations

---

[7] Based on the research literature, this behavior "General Praise" is not counted as an ineffective behavior for grades K-3.

[8] Allows Unison Response is only an ineffective behavior for 4th grade and above.

(elementary and secondary). By far the lowest scoring (most frequently coded) item is General Praise. Other items having scaled scores below 1.95 include, at the elementary level, Multiple Questions Asked as One, Non-Academic Questions and Delays Desist. In secondary classes, other lower scoring items included Allows Unison Response, Multiple Questions Asked As One, Non-Academic Questions and Delays Desist.

---

Insert Figures 4 & 5 About Here

---

Domain 4 Scores Across Subject Areas: Although overall scores for Domain 4 effective items were quite low in Figures 2 and 3, one might generally expect teachers in the sciences or mathematics to exhibit different levels of scores for these items, since their content area training involves proofs as well as inductive and deductive logic. Therefore, subject matter areas were broken into four groups. The breakdown was limited to four general groups due to limitations in the Frame Factor measure used to identify subject areas:

1. Sciences
2. Mathematics
3. Language Arts (including foreign language, history and social studies), and
4. Other (including a broad range of subjects from PE to DCT).

Mean scaled item scores on the Domain 4 items were computed separately for each of these subject areas across elementary and secondary teachers. Figures 6 and 7 show that similar patterns of item performance characterized both elementary and secondary teachers in both years. It appears from the figures that only States Cause and Effect for science teachers and Academic Rules for teachers of mathematics approach the levels found for other FPMS effective items in Figure 3.

At the elementary level, science teachers appear to exhibit higher scores in both the Treatment of Concepts and the Stating of Causes and Effects than do any of the other groups. Regarding Academic Rules and developing Criteria for Value Judgements, however, the only advantage lies with mathematics and language arts teachers for Academic Rules. Also, except for their shared reduced coding relative to science teachers on Treating Concepts

and Stating Causes and Effects, mathematics teachers do not differ from any other group on the other items.

Essentially the same pattern occurs in secondary classes. Specifically, science teachers score higher in both the Treatment of Concepts and the Stating of Causes and Effects than do the other groups, while mathematics teachers exhibit more behaviors relating to Academic Rules.

---

Insert Figures 6 & 7 About Here

---

## Conclusions and Implications

These data show that questioning and interactive behaviors dominate the average lesson taught by teachers in the Florida Beginning Teacher Program. The four effective items in Domain 3: Development represent 41 percent of all behaviors identified in these average lessons. When combined with their ineffective counterparts, they comprise fully half (55% elementary, 50% secondary) of all behaviors exhibited in an average lesson despite representing only 20 percent of all performance categories involved. Certainly the research literature over a fifty-year period supports the use of questioning as an effective behavior, and these data suggest that much of a classroom teacher's effort is seen here. The FPMS summative observation instrument identifies both quality and quantity within five questioning indicators (two effective and three ineffective). Indicator 5C (nonacademic questions) deals with questions including opinion, while 5B questions require justifying, analyzing, etc. from students rather than rote recall.

The most frequently occurring ineffective items (Multiple Questions asked as One, Non-Academic Questions, General Praise - items 5A, 5B, 7) are those that associate with the most frequently occurring effective items (Single Factual Questions, Questions Requiring Analysis/Reasons, and Specific Academic Praise - items 5A, 5B, 7). This is neither unusual nor unexpected, since teachers asking numerous effective questions and giving specific praise during a lesson are more likely to use an ineffective questioning behavior or general praise than those exhibiting fewer such behaviors.

Of greater concern, it appears that the Presentation of Subject Matter items (other than Concept Treatment) only rarely occur during these observations. Across all teachers, these represent only four percent of elementary and seven percent of secondary teacher behaviors. This, despite embodying 20 percent of the performance categories involved. Even among mathematics and science teachers these behaviors tend to be used only in a limited fashion, with the exceptions of Academic Rules in secondary mathematics classes and Concept Treatment in science classes. While the other domains concern important organizational, communication and discipline skills, Domain 4 items document a teacher's management of subject matter. If observations indicate that teachers are not presenting content, one might guess that the frequently occurring questions and responses have to do with learning factual material from texts rather than integrated subject matter. This domain provides evidence that a teacher is working with students in learning content. It is within this domain that the most improvement could occur for Florida's beginning teachers and most probably for teachers throughout the world. This is also the domain in which observers have the most difficulty coding and teachers have the most difficulty practicing. Perhaps because of the perceived need confirmed by these data, the topic of subject matter presentation is currently a very popular area in the educational literature as the plethora of recent articles on higher order thinking and problem solving attest.

Another interesting phenomenon that these analyses brought to light is the relatively rare occurrence of teacher performance in the seatwork/homework indicator. The literature shows that student learning increases if teachers assign seatwork or homework, give clear directions, check student comprehension of assignments and correct and provide feedback on student performance of the assignments (Good & Grouws, 1977; Medley, 1977; Brophy and Evertson, 1976). To the extent that these behaviors are lacking among beginning teachers, students are deprived of this educational asset.

## References


Brophy, J. & Evertson, C. (1976). Learning from Teaching: A Developmental Perspective. Boston: Allyn and Bacon.

Good T, & Grows, D. A. (1977). Teachers manual: Missouri Mathematics Effectiveness Project. Columbia, Missouri. Center for Research in Social Behavior, University of Missouri.

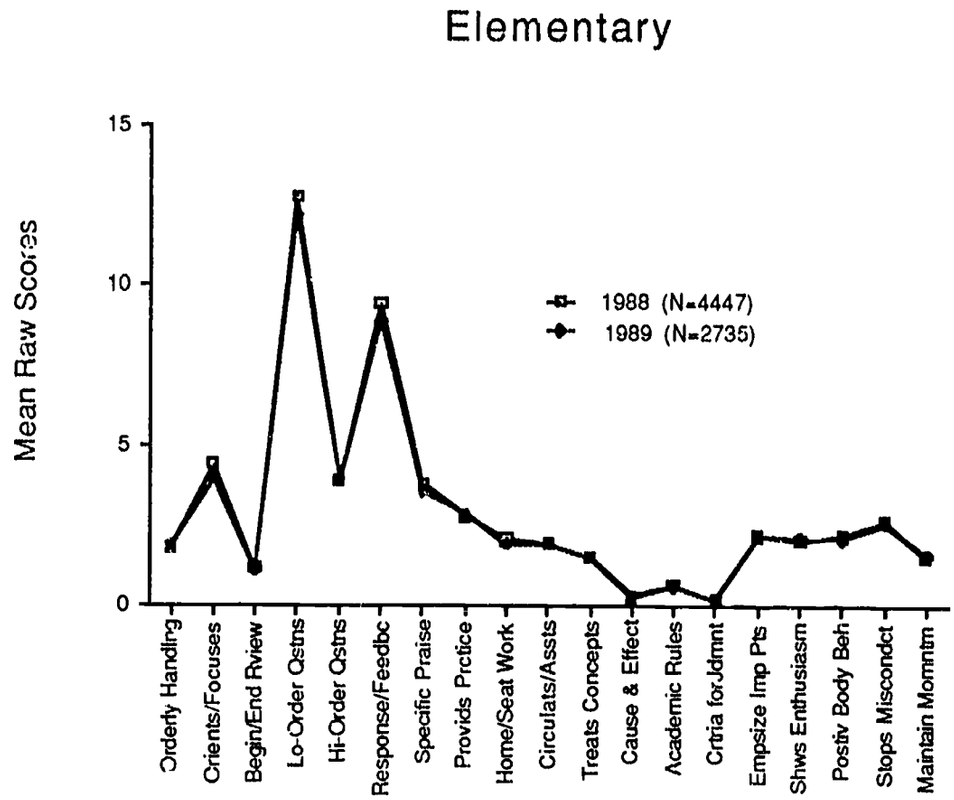
Medley, D. M. (1977). Teacher competence and Teacher Effectiveness: A Review of Process-Product Research. Washington, D.C.: AACTE.

Peterson, D., Micceri, T. & Smith, B. O. (1985). Measurement of Teacher Performance: A Study in Instrument Development. Teacher and Teacher Education, 1:1, 63-77.

Smith, B.O., Peterson, D. & Micceri, T. (1987). Evaluation and Professional Improvement Aspects of the Florida Performance Measurement System. Educational Leadership, 44, p. 16-24.

| EFFECTIVE INDICATORS | INEFFECTIVE INDICATORS |
|---|---|
| **Domain 3 Organization** | |
| 1 Begins Instruction Promptly | 1 Delays |
| 2 Handles Materials in an Orderly Manner | 2 Does Not Organize or Handle Materials Systematically |
| 3 Orients Students to Classwork/Maintains Academic Focus | 3. Allows Talk/Activity Unrelated to Subject |
| 4 Conducts Beginning/Ending Review | |
| 8 Provides for Practice | 8 Extends Discourse, Changes Topic With No Practice |
| 9 Gives Directions/ Assigns/Checks comprehension of Homework, Seatwork Assignments/Gives Feedback | 9 Gives Inadequate Directions on Homework/No Feedback |
| 10 Circulates and Assists Students | 10 Remains at Desk/Circulates Inadequately |
| **Domain 3 Development** | |
| 5A Questions: Single Factual (Domain 5.0) | 5A Allows Unison Response |
| 5B Questions: Requires Analysis/Reasons | 5B Poses Multiple Questions Asked as One |
| | 5C Poses Nonacademic Questions/ Nonacademic Procedural Questions |
| 6 Recognizes Response/Amplifies/Gives Correct Feedback | 6 Ignores Student or Response/ Expresses Sarcasm, Disgust, Harshness |
| 7 Gives Specific Academic Praise | 7 Uses General, Nonspecific Praise |
| **Domain 4 Subject Presentation** | |
| 11 Treats Concepts-Definition/Attributes/Examples/ Nonexamples | 11 Gives Definitions or Examples Only |
| 12 Discusses Cause-Effect/Uses Linking Words/ Applies Law or Principle | 12 Discusses Either Cause or Effect Only/Uses No Linking Word(s) |
| 13 States and Applies Academic Rule | 13 Does not State or Does Not Apply Academic Rule |
| 14 Develops Criteria and Evidence for Value Judgement | 14 States Value Judgement With No Criteria or Evidence |
| **Domain 5 Communication** | |
| 15 Emphasizes Important Points | 17 Uses Vague/Scrambled Discourse |
| 16 Expresses Enthusiasm Verbally/Challenges Students | 18 Uses Loud-Grating, High Pitched, Monotone, Inaudible Talk |
| 19 Uses Body Behavior that Shows Interest-Smiles, Gestures | 19 Frowns, Deadpan or Lethargic |
| **Domain 2 Management of Conduct** | |
| 20 Stops Misconduct | 20 Delays Desist/Doesn't Stop Misconduct/Desists Punitively |
| 21 Maintains Instructional Momentum | 21 Loses Momentum-Fragments, Nonacademic Directions, Overdwells |

Figure 1: Florida Performance Measurement System Indicators



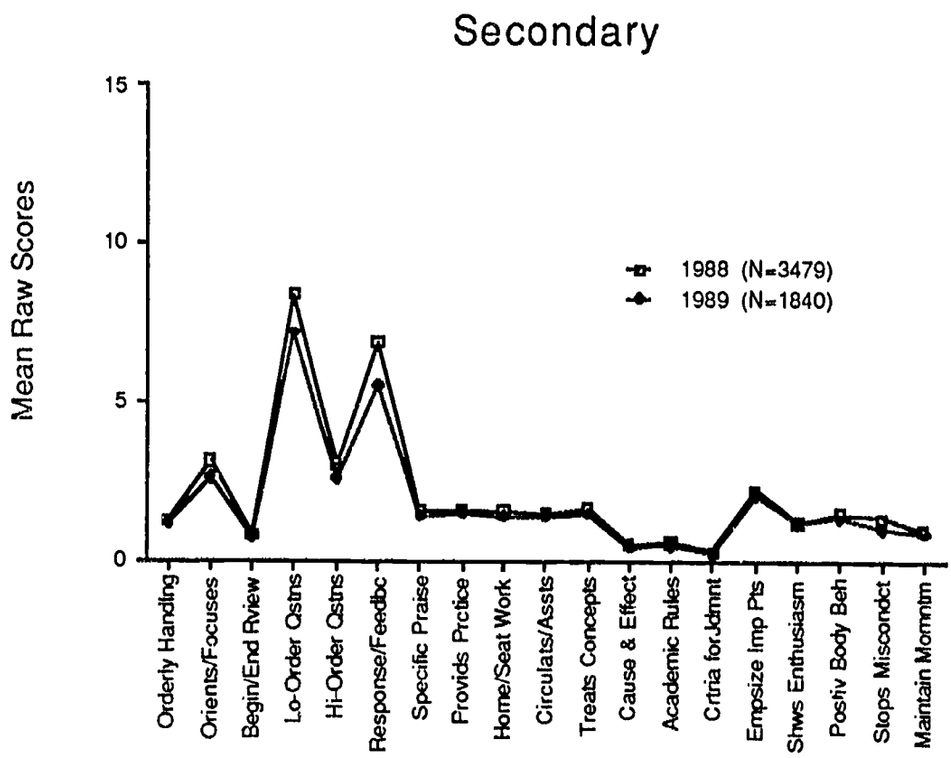
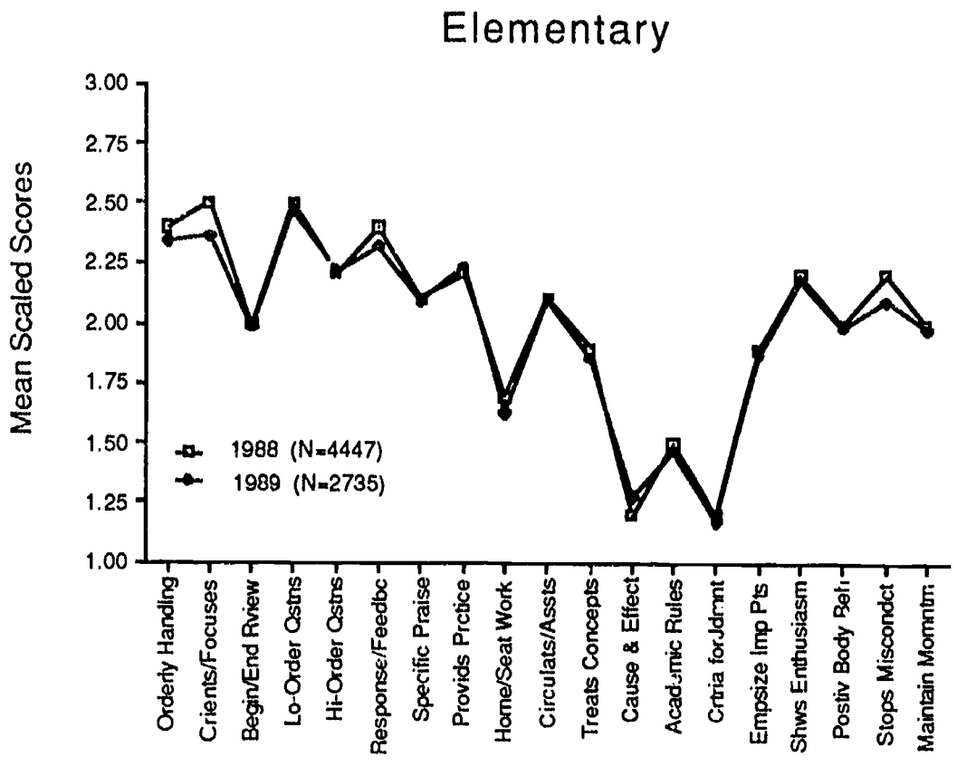




Figure 2: Mean Raw Effective Item Scores of Beginning Teachers During the 1987/88 and 1988/89 School Years Standardized to 30 Minute Lessons.



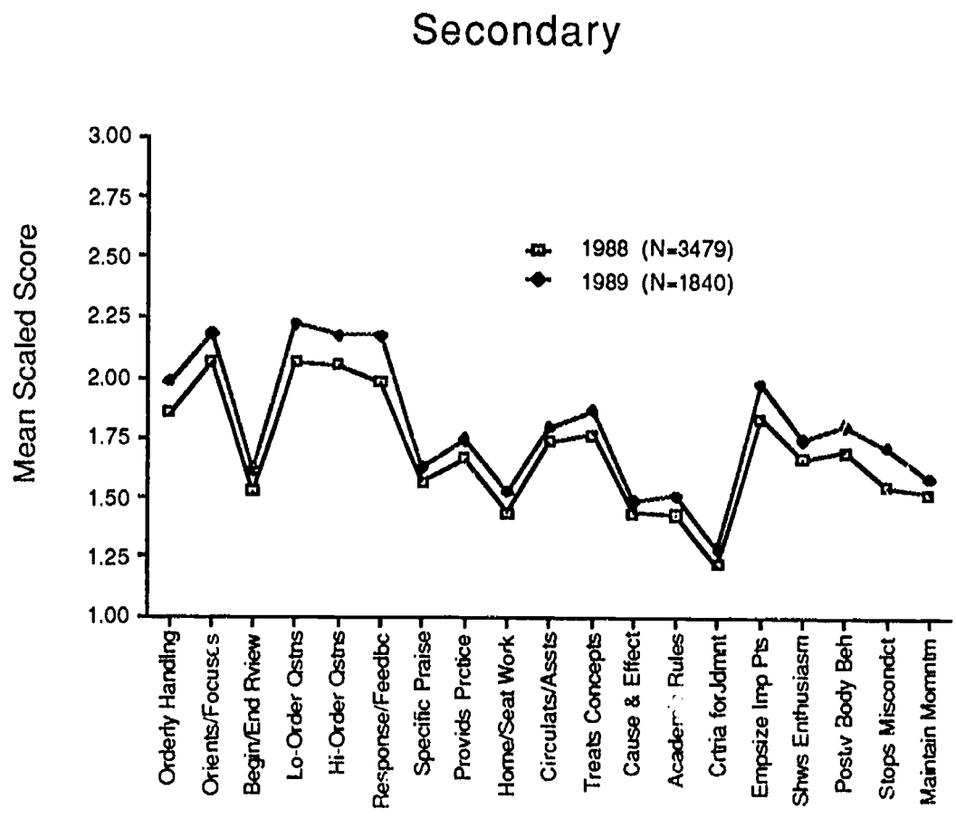
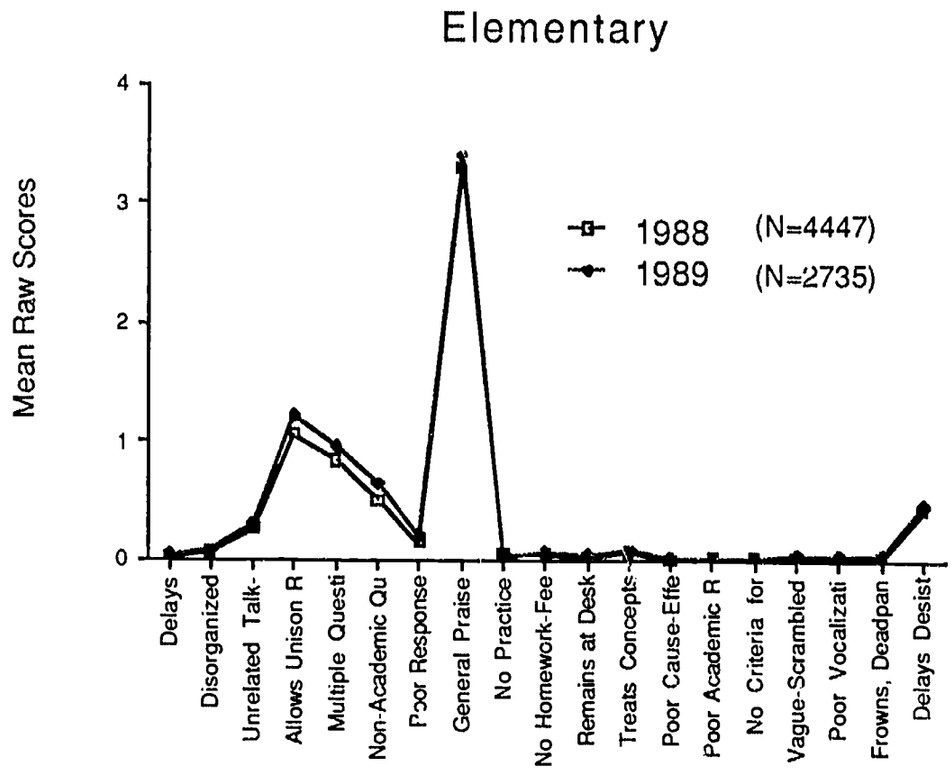




Figure 3: Mean Scaled Effective Item Scores of Beginning Teachers During the 1987/88 and 1988/89 School Years



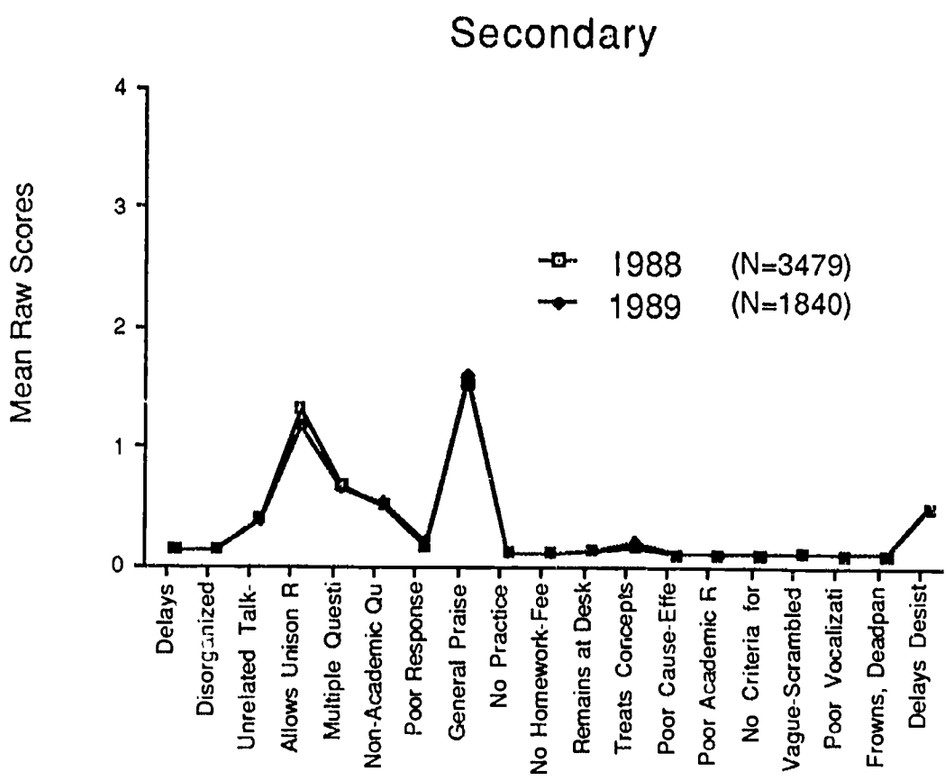




Figure 4: Mean Raw Ineffective Item Scores of Beginning Teachers During the 1987/88 and 1988/89 School Years Standardized to 30 Minute Lessons

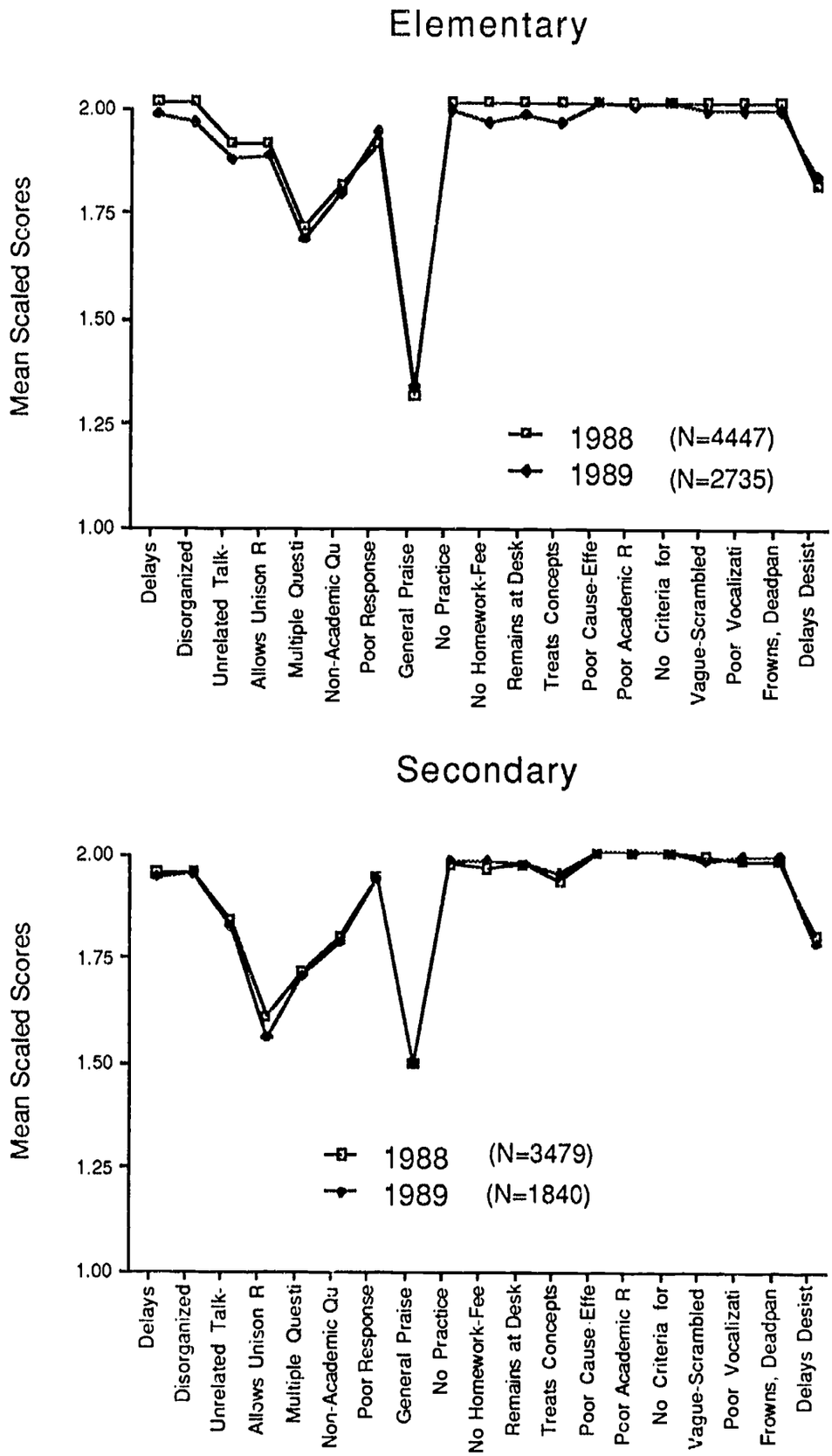



Figure 5: Mean Scaled Ineffective Item Scores of Beginning Teachers During the 1987/88 and 1988/89 School Years

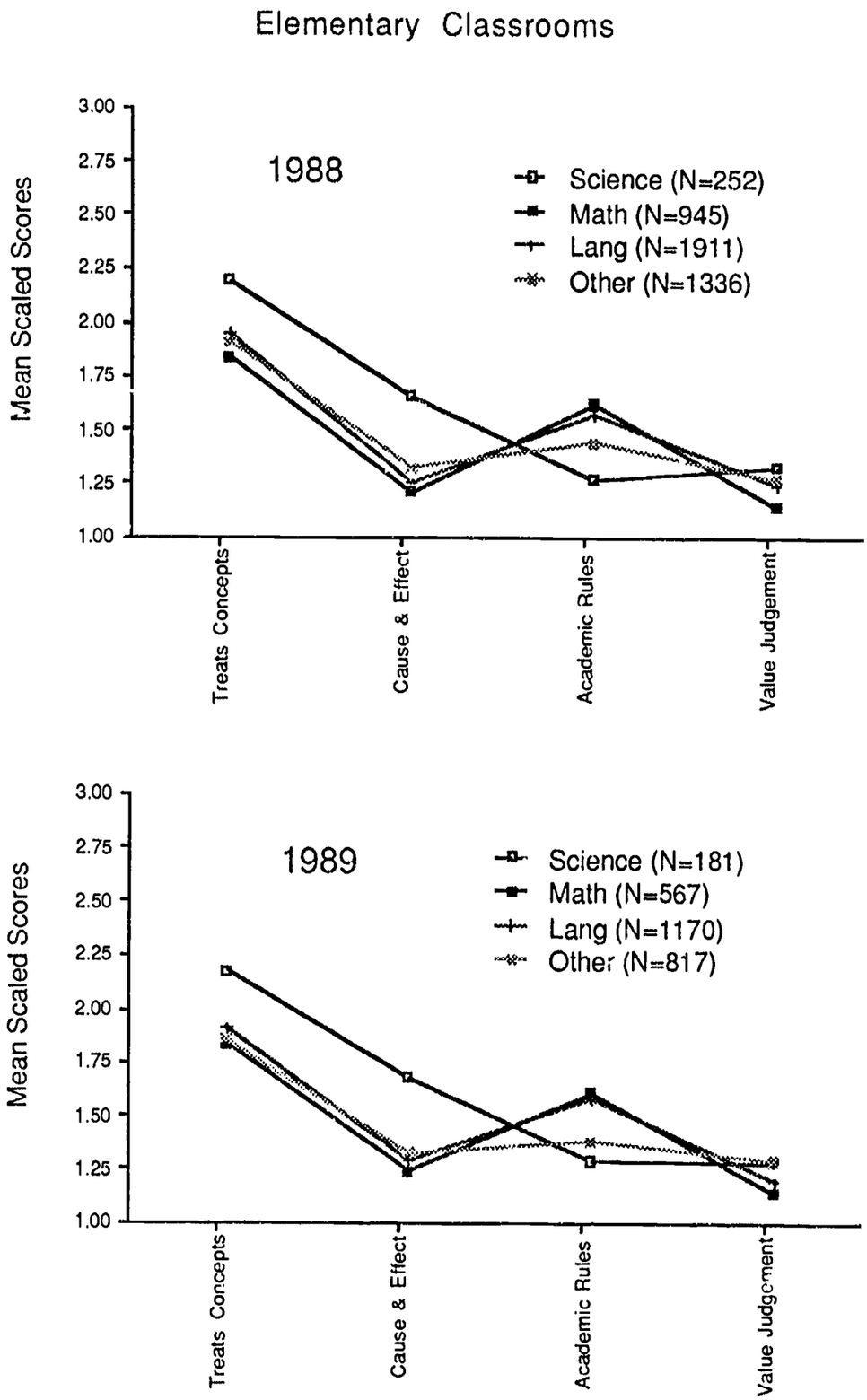
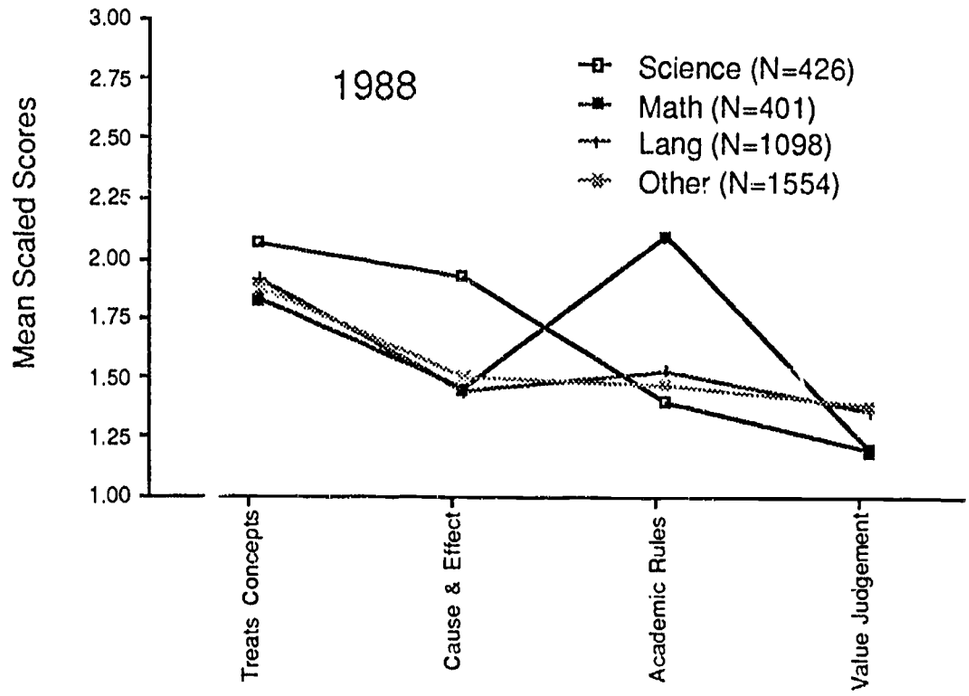



Figure 6: Mean Scaled Effective Domain 4 Scores of Beginning Teachers in Elementary Schools Across Subject Areas

Secondary Classrooms



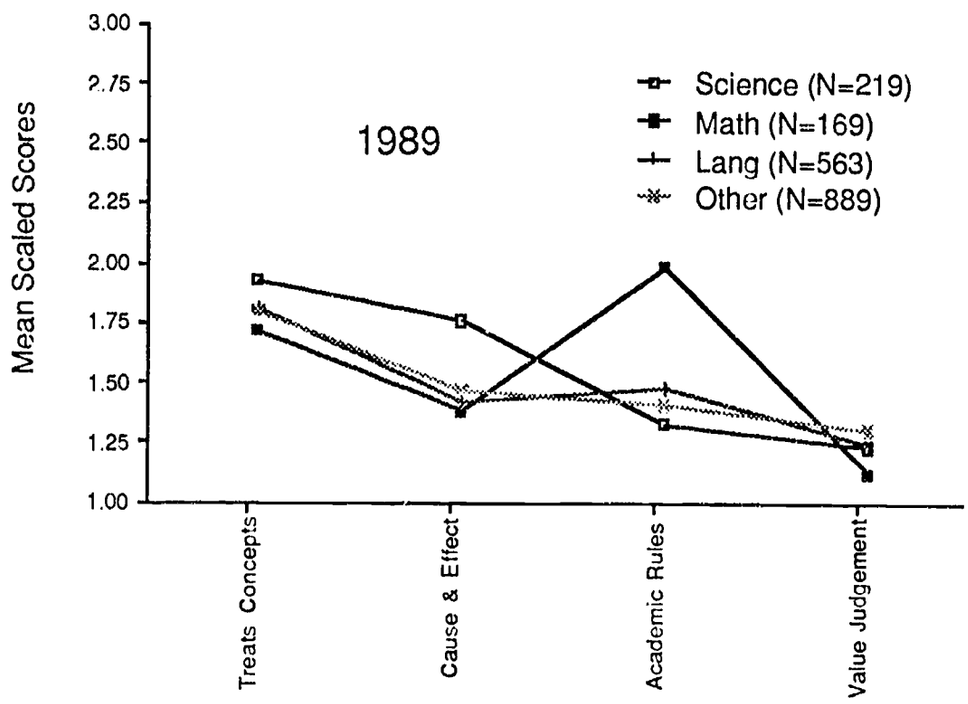




Figure 7: Mean Scaled Effective Domain 4 Scores of Beginning Teachers in Secondary Schools Across Subject Areas