ED 370 789 SE 054 461

AUTHOR Horton, Phillip B.; And Others

TITLE Randomness and Replication Revisited: A Content Analysis of Research Published in "Science Education' from 1988-1992.

PUB DATE Jan 94

NOTE 24p.; Paper presented at the Annual Meeting of the Association for the Education of Teachers in Science (El Paso, TX, January 6-9, 1994).

PUB TYPE Information Analyses (070) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Content Analysis; *Educational Research; Elementary Secondary Education; *Evaluation Methods; Higher Education; Research Utilization; *Science Education; *Theory Practice Relationship

IDENTIFIERS *Review of Educational Research

ABSTRACT

This study was conceived as a systematic replication of a content analysis of published science education research conducted by Horton et al. in 1993. As such, 47 research articles published in "Science Education" between 1988 and 1992 were examined. Also, this study further extended the findings of Shaver and Norton, and Wallen and Fraenkel, who conducted similar analyses of general and social studies research. One major objective in this analysis was to determine whether science education researchers routinely practice commonly recommended research procedures. In addition, reviewers were interested in whether direct or systematic replication, common practices in other disciplines, play significant roles in science education research. The method of analysis and a discussion of the results are included. (ZWH)

RANDOMNESS AND REPLICATION REVISITED: A CONTENT ANALYSIS OF RESEARCH PUBLISHED IN SCIENCE EDUCATION FROM 1988-1992.

Philip B. Horton, Science Education Department,
Florida Institute of Technology,
150 W. University Boulevard, Melbourne, FL 32901-6988

Andrew A. McConney, The Evaluation Center,
Western Michigan University, Kalamazoo, MI 49008-5192

Amanda W. McConney, Department of Science Studies,
Western Michigan University, Kalamazoo, MI 49008-5033

Rodney Bowers
Terri Schoone
Denis Hamelin
Melinda Bier
Joyce Oster
Ed Spencer

Science Education Department, Florida Institute of Technology
150 W. University Boulevard, Melbourne, FL 32901-6988

Paper presented at The Association for the Education of Teachers in Science 1994 Annual Meeting January 6-9 El Paso, Texas.

Abstract

This study was conceived as a systematic replication of a content analysis of published science education research conducted by Horton et al. (1993). As such, we examined 47 research articles published in Science Education between 1988 and 1992. Also, this study further extended the findings of Shaver and Norton (1980a; 1980b) and Wallen and Fraenkel (1988a; 1988b) who conducted similar analyses of general and social studies education research. As in our previous work (Horton et al., 1993), we sought to determine whether science education researchers routinely practice (as evidenced by published work) commonly recommended research procedures (e.g., Ary, Jacobs, & Razavieh, 1985; Borg & Gall, 1989; Campbell & Stanley, 1963; Fraenkel & Wallen, 1990). We were also interested in whether direct or systematic replication, common practices in other disciplines, play significant roles in science education research.

This study's findings indicate marginal improvements in some areas for the science education research analyzed as compared to the previous five-year period reviewed in the Journal of Research in Science Teaching (Horton et al., 1993). However, our findings also indicate that few of the published studies provide adequate descriptions of populations (26%) or samples (17%), and fewer still employ random selection (6%) or assignment (9%) as methods of equating groups. Yet, we found that in about half of the articles, authors' conclusions were not suitably limited by the populations and samples studied. These problems are further exacerbated by the persistent lack of direct or systematic replication (only 2% of the studies reviewed) in published science education research (methods which could mitigate limited generalizability). Given these findings, we are not surprised by the seemingly common perception at the public and policy levels that educational research does not seriously inform science classroom practice (Kaestle, 1993; Wright, 1993). We would argue however, that it is not the agricultural and psychological research models that are necessarily lacking. Rather, we suggest that researchers, reviewers, and interested readers pay far greater attention to the fundamental, underlying assumptions of these models lest the "baby be thrown out with the bath-water".

Introduction

For the education community in general and science education researchers in particular, discussion on the relevancy of educational research in guiding and informing ciassroom practice continues (Wright, 1993; Kaestle, 1993; Shymansky & Kyle, 1992; Yeany, 1991). As Wright (1993), then president of the National Association for Research in Science Teaching described, an often heard criticism of educational research is that "simplistic agricultural and psychological models of research" continue to be applied to answer complex, multivariate, multidimensional questions (p. 2). However, another way of viewing this criticism could be to rigorously question whether or not the underlying assumptions and limitations inherent in these models are generally recognized and adhered to by educational researchers. Further, as Wright also points out, we must ask whether direct or systematic replication of research plays the significant role it should--as it does in other scientific disciplines--in building the educational findings and theories on which science classroom practice c ıuld be based.

One way in which these questions can be answered is through content analyses of published educational research. Shaver and Norton conducted two such analyses in which they examined research published in the American Educational Research Journal (AERJ) (1980a) and in two social studies journals Theory and Research in Social Education (TRSE) and Social Education (SoE) (1980b). In both cases, Shaver and Norton found that the underlying and fundamental assumptions (random selection/assignment from adequately described populations) of recommended research methods were not found in much research in social education. These researchers also found that direct or systematic replication were rarely used strategies by which the generalizability of educational research findings could be strengthened or supported. In follow-up content analyses of published social studies research, Wallen and Fraenkel (1988a, 1988b) found the situation described by Shaver and Norton essentially unchanged.

Most recently, Horton et al. (1993) asked similar questions of science education research published in the Journal of Research in Science Teaching (JRST) from 1985 - 1989 (then the most recent five-year period). Similar to the four previous content analyses cited, this study's results showed that because of a persistent lack of randomization and adequate sample and population description, appropriate generalizations beyond the confines of the reported studies may be impossible for most (64%) JRST studies surveyed. Horton and colleagues also found that less than half (48%) of the science education studies properly restricted their conclusions based on the limits imposed by the accessible populations and samples used, and that replication was an even more rarely used strategy (3% of the 130 studies) in science education research than in social education.

In the interests of extending these findings for published science education research, to add to the continuing discussion on the relevancy of educational research, to present a fair case for recommended "agricultural and psychological models" so that these are not prematurely discarded, and to inform the community and readership of the journal, this study sought to ask many of the same questions (and some new ones) of published research in Science Education (SE).

Questions

The first five questions posed in this study are a direct replication of those asked by Horton et al. (1993) while the last five examine in more detail four other areas which could be of interest to the science education research community, and compare this study's results with those previously reported:

(1) Do science education researchers typically select their samples randomly from defined and/or described accessible populations?

(2) Do science education researchers typically define their target populations and describe their samples?

(3) Is replication a frequently used research strategy in science education?

(4) Do science education researchers typically restrict their conclusions based on the limitations of their sampling techniques, or in regard to possible differences between their accessible and target populations?

(5) In consideration of possible threats to validity, do science education researchers typically provide alternative explanations for positive findings?

(6) Is a relationship between research strategy and validity threats evident in this sample of published science education research?

(7) Is a relationship between research strategy and the qualification of study conclusions evident in this sample of published science education research?

(8) Is a relationship between the reliability and validity of assessment tools and statistically significant study results evident in this sample of published science education research?

(9) Is a relationship between sample size and statistically significant study results evident in this sample of published science education research?

(10) How do the results for this analysis of SE compare with our findings for JRST?

Method

The rating team for this content analysis comprised one faculty member of the Science Education department (the senior author) and eight science education doctoral students. The senior author has taught graduate courses in statistics and research design for 15 years, and as a prerequisite to participation in this project, the eight students had all successfully completed graduate courses in statistics and educational research design.

In accordance with Krippendorf's (1980) recommendation for the content analysis of a periodic publication and to give us a representative idea of the status of research over the most recent five-year period (data were gathered in fall 1992 and winter 1993), we used purposive selection to choose the volumes of SE (volumes 72 - 76) analyzed. Purposive sampling insures a balanced coverage of the time period of interest whereas random sampling might result in a predominance of data from a more restricted time range. From this five-year (1988 – 1992) pool of volumes we randomly selected issues 72(1), 73(1, 2, 4), 74(2, 4), 75(2, 5), and 76(2, 4, 5) for analysis. We emphasize that this study was descriptive in nature and as such our selection procedure did not initially employ random sampling. Therefore, broad generalizations to the population of published science education research studies outside of the eleven SE issues analyzed were avoided.

A total of 47 research studies were reported in the eleven SE issues examined; these were used as the sample in this study. Because of the nature of our study we did not include articles which fell into the following categories: literature reviews or meta–analyses, instrument development or validation studies, philosophical inquiries, position papers, or historical studies. The breakdown of the types of studies we reviewed is given in Table 1.

Insert Table 1 here

The classification shown in Table 1 was based on Campbell and Stanley's (1963) taxonomy:

1. Pre-experiments: these studies employed no control group and no random assignment. They typically involved a one-group pretest-posttest design, or the static group comparison.

2. True experiments: subjects were randomly assigned to control and treatment groups. The treatment was under the control of the researcher.

3. Quasi-experiments: these studies were similar to true experiments but without random assignment to groups.

4. Correlational: measures on a group of subjects were correlated.

5. Survey: data were gathered using a written questionnaire or oral interview. No treatment was administered.

6. Causal-comparative (ex post facto): groups already different on some independent variable were compared on a dependent variable.

7. Ethnographic: the daily activities of subject(s) in naturalistic settings were detailed.

To collect this study's data we employed a modified coding instrument based on commonly accepted standards of research (Ary, Jacobs, & Razavieh, 1985; Borg & Gall, 1989; Fraenkel & Wallen, 1990). This instrument had been previously used by Horton et al. (1993) for their content analysis of JRST and all of the categories used in the previous study were included as they were also applicable to this study's focus. However, as a result of our initial discussions a number of new categories of interest to the rating team were developed and added to the instrument.

To refine and clarify our coding instrument and as training for the actual study, the team read and rated a number of articles from earlier volumes of SE. When we used the final form of the instrument to rate two articles we calculated Spearman reliability coefficients ranging from .83 to .91 among four pairs of raters. We also used Scott's pi (Krippendorf, 1980) to measure the reproducibility of raters' coding. Scott's pi takes into account the percentage of coded items which would be expected to agree merely by chance. When applied, we obtained reliability estimates ranging from .59 to .94 (average pi = .84) for 12 pair-wise comparisons for the two articles. These reliability coefficients represent average coding agreement 84% higher than what one would expect due to chance.

The 47 studies were read and analyzed over a 6-month period. At least two of the study's authors read and rated each article. Readers rated their articles separately and then met with a team member(s) to develop the final coding for each article. The team met weekly to discuss and develop consensus on areas of disagreement. We also periodically rated a common article and compared results to insure that we were consistently using the same procedures. Finally, all data were jointly transferred to a spreadsheet by two of the study authors. Any coding anomalies were noted and brought to the team for clarification to insure consistency with agreed–upon definitions.

Results and Discussion

As shown in Table 1, the majority (38%) of the 47 SE articles analyzed were surveys. Only about one-third of the studies (30%) had some type of experimental design (as compared to 52% of JRST articles). Ethnographic research accounted for one-fifth of the 47 studies analyzed, continuing the upward trend for this methodological appioach to educational research seen in our analysis of JRST (Horton et al., 1993).

Table 2 summarizes SE study authors' reasons for conducting their research. As was the case for AERJ, TRSE, and JRST, the majority of SE authors in these eleven issues justified their work by either explicit or implicit arguments of worth (58%). Also similar to our JRST analysis, tests of theory (9%) and replication studies (2%) were noticeably few among the articles coded. We find it troubling yet revealing that direct or systematic replication studies should continue to be so rare in a representative sample of published research despite many calls to the contrary by the science education academic community (Shymansky & Kyle, 1992). More encouraging however, was that the reported purpose of 23% of the 47 studies was to extend previous research findings. This is a substantial increase over the 5% seen in JRST, but nothing like the 80% Shaver and Norton (1980a) reported for AERJ. It may be that Shaver and Norton included authors' arguments for a study's worth (explicit and implicit) in this category.

Insert Table 2 here

Table 3 provides this study's findings on the sampling and group formation protocols for the 47 SE studies analyzed. Substantially more population and sample descriptions were given by authors in SE than in JRST. Despite this, and as previously found for JRST and TRSE, almost two-thirds of the sample descriptions were deemed "marginal" while samples formed by convenience (typically intact groups readily available to authors) remained by far the most common (68%). Random selection (6%) and assignment (3%), or the random assignment of treatments to intact groups (6%) were little used procedures in these 47 studies. Most significantly, randomization was a part of the group formation protocol in only 3 of the 14 SE articles in which some type of experimental design was employed. Coupled with the lack of accessible population description in 75% of the studies and absent or marginal sample descriptions in 80%, this would seem to indicate that generalizations beyond the limits of particular projects are unlikely or impossible for most of the 47 SE articles analyzed.

Insert Table 3 here

The breakdown of SE authors' qualification of their research conclusions is given in Table 4. Again, we noted an improvement in the percentage of properly qualified conclusions (45%) as compared to JRST (35%) and TRSE (31%). However, it was also evident that overgeneralization (Raths, 1973) remained a common practice in the published science education research examined here.

Insert Table 4 here

As shown in Table 5, instrumentation (inadequate demonstration of the reliability/validity of assessment tools) and selection bias (pre-existing differences in subject characteristics may account for the results) were the validity threats most common to the SE studies analyzed. About one-half of the 47 studies were also threatened by the effects of history, mortality, and novelty. Despite this, we found that in only 17% of the studies were validity threats acknowledged and discussed--and then only in a marginal fashion. None of the 47 SE articles were coded as having satisfactorily discussed threats to validity. These results were similar to those found for JRST, and leave readers with unanswered questions as to possible alternative explanations for the conclusions reached in many of these 47 SE articles.

Insert Table 5 here

Table 6 relates threats to internal validity with experimental design. As evident, all of the research designs seen in this content analysis were subject to at least one threat. The one true experimental study was threatened by mortality only, while at the other extreme pre-experiments were typically subject to a number of validity concerns, most likely pretest sensitization and history. Quasi-experimental, causal comparative (ex post facto), and ethnographic studies were also typically subject to multiple threats--usually selection bias and instrumentation. Ex post facto studies also seemed particularly subject to history threats. Also noteworthy from Table 6 was that (with the possible exception of causal comparative studies and disregarding the one true experiment and one content analysis) the lack of alternate explanations noted above was common in all research designs.

Insert Table 6 here

In Table 7 we examined the extent to which SE authors' qualification of study conclusions were related to particular research designs. As evident from the results in Table 7 (and disregarding the one true experiment and one content analysis), the authors of survey studies were most likely to offer properly qualified conclusions justified by data (72%). In no other design did we find more than 50% of the studies to offer conclusions properly qualified by limitations in sampling or design protocols. All designs suffered from improperly qualified conclusions, with overgeneralization being most common in ex post facto (83%) and ethnographic (50%) research (again disregarding the one true experiment and one content analysis). The style of reporting conclusions as findings was most common in quasi-experimental designs.

Insert Table 7 here

We were also interested in determining whether there was a relationship between studies reporting statistically significant results and the reliability/validity of their measur'ng instruments. Figure 1 illustrates this relationship for the 29 SE studies reporting significant results. No trend was evident from this figure; that is, statistically significant results did not seem to depend on reported reliable or valid measuring instruments. Of course it is entirely possible that many of the authors of studies we coded as having inadequate or unreported reliability/validity of assessment tools simply chose to omit these from their articles.

Insert Figure 1 here

Similarly, we were curious to examine the relationship between sample size and the presence of statistically significant results for the 29 studies. Figure 2 illustrates this relationship. Again, no clear trend is shown by the figure. However, it was noted that larger sample sizes did not necessarily guarantee statistical significance. Nor are statistically significant results the raison d'être for experimental research. This is an important point for researchers who shy away from experimental research designs or inferential statistics because of the twin misconceptions that very large sample sizes are needed for experiments, and/or that non significant results are unacceptable in educational research.

Insert Figure 2 here

Limitations and Conclusions

This content analysis was conceived as a descriptive study of research and reporting practices commonly found over a five-year period in the journal Science Education We therefore hasten to remind readers that what we are providing is a static representation of the state of the art in one science education research journal. One metaphor that may be useful is the motion picture metaphor. This analysis provides viewers with five partial still frames of a motion picture about eighty frames long. Just over seventy frames came before our still pictures, and a couple came after. Similar to the point we press for educational research, no one frame, research method, or study tells the whole story. However, viewers and researchers can be informed and their practice guided by the five frames. We offer the following conclusions in this spirit.

In answer to question 1, we conclude that randomization protocols were rarely used by science education researchers in the 47 articles coded. As only about 30% of the articles were experimental, this may seem unimportant. However, when combined with the fact that about three-quarters of the studies lacked adequate population or sample descriptions (question 2), we are left to conclude that at least 75% of the articles' authors have little justification for generalizing beyond the limits of their particular research.

Compounding this want of justifiable generalizability in the 47 studies, and in answer to question 3, was a notable absence of direct or systematic replication as a research strategy. As Carver (1978) points out, replication can be one solution to the limited generalizability found in much educational research. We were encouraged by the 23% of studies which were coded as extending findings.

The answer to question 4 was also no: over half of the researchers reporting these 47 studies did not restrict or qualify their conclusions based on limitations in sampling technique or study design. Overgeneralization remained a common problem for many science education researchers, particularly for those engaged in causal comparative or ethnographic studies (question 7). Similarly, despite the multiple threats to the validity of all designs found in these 47 studies (question 6), alternative explanations for research findings were marginally entertained (question 5) by very few authors (17%).

No clear trends emerged from our data in answer to questions 8 and 9. Authors reporting statistically significant results were just as likely to do so whether or not their assessment instruments were coded as adequately reliable or valid. It is likely that many authors simply omit this important information from their studies--an unfortunate practice as this may be seen as another barrier to replication. Similarly, for the 29 studies using inferential statistics, no clear trend emerged relating sample size to statistically significant results.

Finally, in comparison to our results for JRST (1985 - 1989), the results of this analysis of SE show marginal improvements in some areas, and a decided shift away from experimental research to survey and ethnographic studies. Areas of improvement included more adequately described populations and samples, and more properly qualified conclusions. However a number of fundamental methodological (lack of randomization, replication) and reporting (lack of population/sample descriptions, lack of reliability/validity evidence, lack of alternative explanations, overgeneralization of conclusions) problems remain. The paucity of replication studies in a discipline highly variable in nature is puzzling and perhaps the most pressing problem we need to address as a scientific community.

It was therefore evident from the answers to these questions that simplistic models of research are not necessarily at fault for the public or policy-makers' perception that educational research does little to inform practice. It seems as plausible to suggest that the users of these models are either unaware or choose to ignore the basic assumptions inherent in the models. The evidence for this lies in the practices found common in these 47 SE articles. This is not to say that all or even most research should be conducted in highly controlled laboratory-like environments divorced from actual science classrooms. Rather, we would hope that all available models with explanatory power would be used, with the caveat that the limitations inherent in these models be recognized. We therefore urge a greater awareness of the need for fully described and reported study variables such as population and sample characteristics, so that interested researchers, users, and policy-makers can make more informed judgments on the relevancy of research to practice in the science classroom.

References

Ary, D., Jacobs, L. C., & Razavieh, A. (1985) Introduction to research in education (rev. ed.). New York: Holt, Rinehart and Winston.

Borg, W. R. & Gall, M. D. (1989). Educational research (rev. ed.). New York: Longman.

Campbell, D., & Stanley, J. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48(3), 378-399.

Fraenkel, J. R. & Wallen, N. E. (1990). How to design and evaluate research in education. New York: McGraw-Hill.

Horton, P. B., McConney, A. A., Woods, A. L., Barry, K., Krout, H.L., Doyle, B.K. (1993). A content analysis of research published in the Journal Of Research In Science Teaching from 1985 - 1989. Journal of Research in Science Teaching 30(8), 857-870.

Kaestle, C. F. (1993). The awful reputation of education research. Educational Researcher, 22(1), 23-31.

Raths, J. (1973). The emperor's clothes phenomenon in science education. Journal of Research in Science Teaching, 10, 201-211.

Shaver, J. & Norton, R. (1980a). Randomness and replication in ten years of The American Education Research Journal. Educational Researcher, 44(3), 265-291.

Shaver, J. & Norton, R. (1980b). Populations, samples, randomness, and replication in two social studies journals. Theory and Research in Social Education, 8(2), 1-10.

Shymansky, J. A. & Kyle, W. C. Jr. (1992). Establishing a research agenda: Critical issues of science curriculum reform. Journal of Research in Science Teaching, 29, 749-778.

Wallen, N. & Fraenkel, J. (1988a). An analysis of social studies research over an eight year period. Theory and Research in Social Education, 16(1), 1-22.

Wallen, N., & Fraenkel, J. (1988b). Toward improving research in social studies education. Boulder, CO: Social Science Education Consortium.

Wright, E. L. (1993). The irrelevancy of science education research: Perception or reality? NARST News, 35(1), 1-2.

Yeany, R. H. (1991). Dissemination and implementation of research findings: Impacting practice. NARST News, 33(4), 1.

Table 1

A Sampling of the Types of Studies Found in SE from 1988 - 92 and JRST from 1985 - 89

Study Type	JRST	SE
Pre-experimental	8%	9%
True experimental	24%	2%
Quasi-experimental	20%	19%
Correlational	12%	0%
Survey	12%	38%
Causal comparative	24%	15%
Ethnographic	10%	21%
Content Analysis	NA	2%

Note. Both content analyses studies may have had more than one design so that percentages do not sum to 100%. NA = data not reported. N = 47 for SE; N = 130 for JRST.

Table 2

Author's Purposes for Studies Analyzed in AERJ, TRSE, JRST, and SE

Author's Purpose	AERJ (n=151)	TRSE/1988 (n=46)	JRST (n=130)	SE (n=47)
Test of theory	8%	NA	5%	9%
Implied test of theory	NA	NA	NA	6%
Replication	14%	NA	3%	2%
Extend previous findings	80%	NA	5%	23%
Explicit argument of worth	NA	76%	72%	43%
Worth implied	NA	20%	18%	15%
Other	4%	NA	0%	0%
None given	6%	4%	0%	2%

Note. Some articles included multiple studies so that section totals may exceed 100%.
NA = data not reported.

Table 3

Sampling and Group Formation Protocols for AERJ, TRSE, JRST, and SE

Protocols	AERJ (n=151)	TRSE1988 (n=46)	JRST (n=130)	SE (n=47)
Population				
Target defined	32%	31%	5%	19%
Accessible described	8%	NA	12%	26%
Sampling procedure				
Random selection	19%	4%	15%	6%
Volunteers	9%	8%	12%	13%
Convenience	12%	62%	62%	68%
Claimed representative	9%	13%	4%	2%
Can't tell	54%	13%	7%	13%
Sample description				
Adequate description	NA	17%	8%	17%
Marginal description	85%	63%	57%	62%
No description	15%	20%	35%	21%
Group formation				
Random assignment to groups	35%	15%	24%	3%
Treatment randomly assigned	NA	NA	12%	6%

Note. Some articles included multiple studies so that section totals may exceed 100%. NA = data not reported.

'able 4

Authors' Qualification of Conclusions for TRSE, JRST, and SE

Catcgorics	TRSE/1988 (n=46)	JRST (n=130)	SE (n=47)
Properly Qualified			
Data justify conclusions	NA	48%	57%
Qualificd properly	31%	35%	45%
Improperly Qualified			
Reported as truth	NA	15%	11%
Reported as trivia	NA	0%	0%
Reported as findings	NA	8%	9%
Overgeneralizes	48%	48%	43%

Note. Section totals may exceed 100% as assignment of study conclusions to more than one category was possible.

NA = data not reported.

Table 5

Threats to Validity for Studies in TRSE, JRST, and SE

Threats to validity	TRSE/1988 (n=46)	JRST (n=130)	SE (n=47)
Internal validity			
Order	NA	6%	2%
Regression	0%	17%	32%
Maturation	0%	21%	23%
Instrumentation	2%	52%	70%
Pretest	4%	29%	32%
History	4%	38%	53%
Mortality	21%	47%	49%
Selection bias	32%	53%	85%
Ecological validity			
Novelty/disruption effect	NA	NA	47%
Multiple treatment interference	NA	NA	6%[a]
Treatment not explicit	NA	NA	38%[a]
Experimenter effect	42%	56%	60%
Hawthorne or John Henry effect	15%	48%	56%[a]
Threats discussed	21%	24%	17%
Satisfactorily	NA	5%	0%
Marginally	NA	19%	17%

Note. Some articles included multiple studies so that section totals may exceed 100%. NA = data not reported. [a]These percentages were calculated on the 16 studies that reported a treatment.

Table 6

Research design and threats to internal validity for 47 SE studies

	Pre-	True	Quasi-	Survey	Ex post	Ethno-	Content
	expt'l.	expt'l.	expt'l.		facto	graphic	analysis
	(n=4)	(n=1)	(n=9)	(n=18)	(n=6)	(n=10)	(n=1)
Threats							
Order	0%	0%	0%	0%	16%	0%	0%
Regression	50%	0%	44%	18%	50%	20%	100%
Maturation	25%	0%	11%	18%	25%	40%	0%
Instrumentation	50%	0%	56%	71%	100%	100%	100%
Pretest	100%	0%	56%	12%	50%	20%	0%
History	100%	0%	44%	29%	100%	70%	100%
Mortality	25%	100%	67%	59%	67%	20%	0%
Selection bias	75%	0%	78%	82%	100%	100%	100%
Average threats per study	5	NA	4	3	4	4	NA
Alternative explanations offered by author(s)	25%	100%	11%	6%	83%	0%	100%

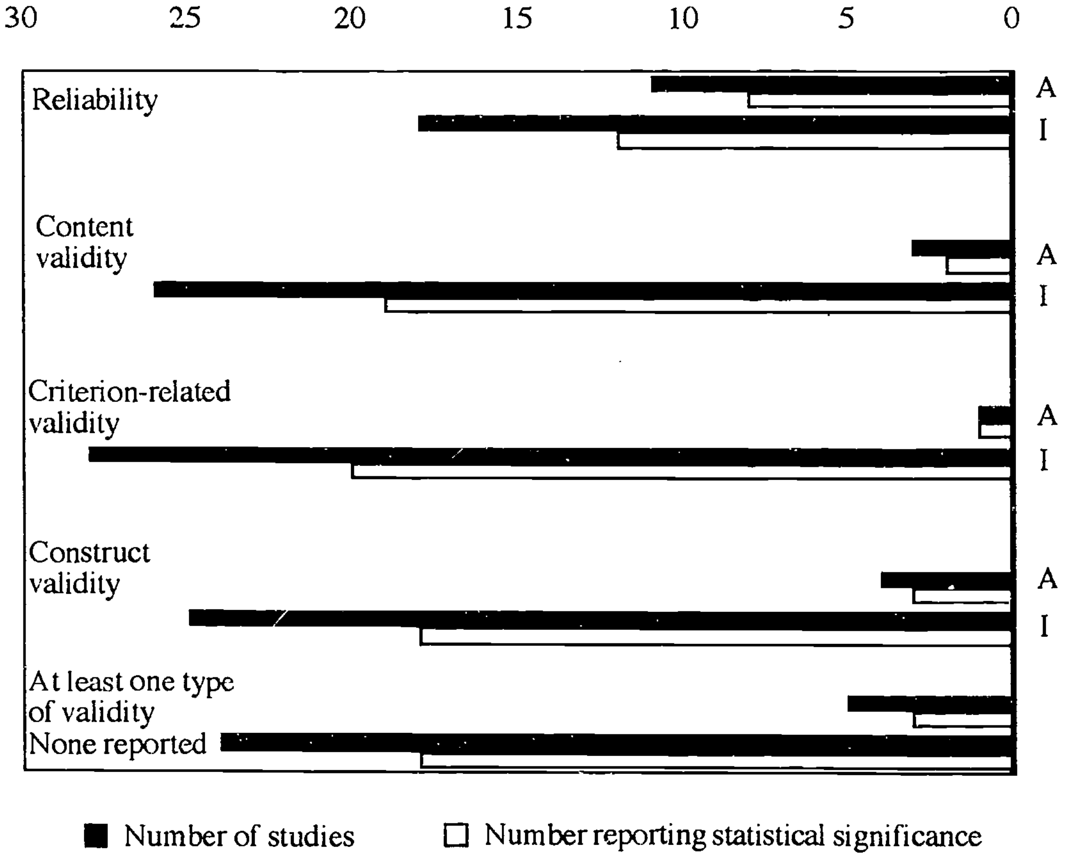
Note. N = 49 as two studies had multiple designs. NA = not applicable.

Table 7

Research design and the Qualification of Conclusions for 47 SE studies

	Pre- expt'l. (n=4)	True expt'l. (n=1)[a]	Quasi- expt'l. (n=9)	Survey (n=18)	Ex post facto (n=6)	Ethno- graphic (n=10)	Content analysis (n=1)[a]
Categories							
Data justify conclusions	75%	0%	56%	72%	50%	40%	0%
Qualified properly	50%	0%	44%	72%	33%	30%	0%
Reported as truth	25%	0%	11%	0%	33%	0%	0%
Reported as trivia	0%	0%	0%	0%	0%	0%	0%
Reported as findings	0%	0%	44%	6%	17%	20%	0%
Overgeneralizes	25%	100%	0%	22%	83%	50%	100%

Note. N = 49 as two studies had multiple designs. [a]Due to a sample size of one, these studies are excluded from the discussion of results.



A = adequate; I = inadequate or not reported

Figure 1. Reliability/validity of assessment and statistically significant results for 29 SE studies.

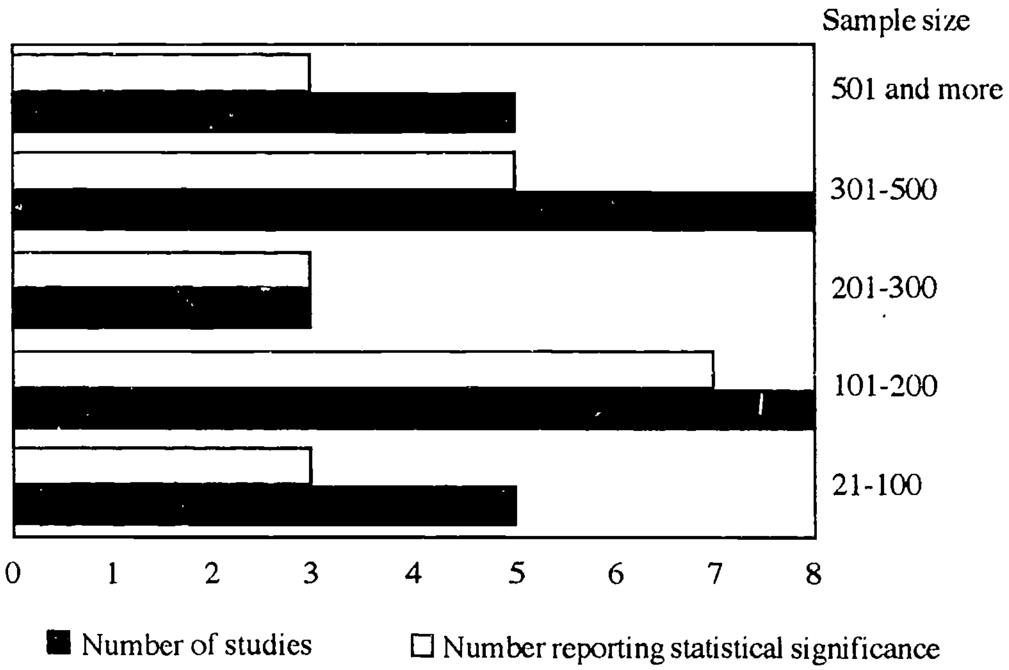


Figure 2. Sample size and statistically significant results for 29 studies in SE.